Received: 5 August 2020 | Revised: 24 March 2021 | Accepted: 12 April 2021

DOI: 10.1111/biom.13471

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Semiparametric estimation of structural nested mean models with irregularly spaced longitudinal observations

Shu Yang

Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA

Correspondence
Shu Yang, Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, NC, USA.
Email: syang24@ncsu.edu

Funding information
National Institute of Environmental Health Sciences, Grant/Award Number: 1R01ES031651; National Cancer Institute, Grant/Award Number: P01 CA142538; Division of Mathematical Sciences, Grant/Award Number: 1811245; National Institute on Aging, Grant/Award Number: 1R01AG066883

Abstract

Structural nested mean models (SNMMs) are useful for causal inference of treatment effects in longitudinal observational studies. Most existing works assume that the data are collected at prefixed time points for all subjects, which, however, may be restrictive in practice. To deal with irregularly spaced observations, we assume a class of continuous-time SNMMs and a martingale condition of no unmeasured confounding (NUC) to identify the causal parameters. We develop the semiparametric efficiency theory and locally efficient estimators for continuous-time SNMMs. This task is nontrivial due to the restrictions from the NUC assumption imposed on the SNMM parameter. In the presence of ignorable censoring, we show that the complete-case estimator is optimal among a class of weighting estimators including the inverse probability of censoring weighting estimator, and it achieves a double robustness feature in that it is consistent if at least one of the models for the potential outcome mean function and the treatment process is correctly specified. The new framework allows us to conduct causal analysis respecting the underlying continuous-time nature of data processes. The simulation study shows that the proposed estimator outperforms existing approaches. We estimate the effect of time to initiate highly active antiretroviral therapy on the CD4 count at year 2 from the observational Acute Infection and Early Disease Research Program database.

KEYWORDS
causality, counting process, discretization, *g*-estimator, martingale

1 | INTRODUCTION

The gold standard to draw causal inference of treatment effects is designing randomized experiments. However, randomized experiments are not always feasible due to practical constraints or ethical issues. In these cases, observational studies are useful. In observational studies, confounding presents a unique challenge to drawing valid causal inferences of treatment effects. For example, sicker patients are more likely to take the active treatment, whereas healthier patients are more likely to take the control treatment. Consequently, it is not fair to compare the outcome from the treated group and the control group directly. Moreover, in longitudinal observational studies, confounding is likely to be time-dependent, in the sense that time-varying prognostic factors of the outcome affect the treatment assignment at each time, and thereby distort the association between treatment and outcome over time.

In these cases, traditional regression methods are biased even adjusting for the time-varying confounders (Robins and Hernán, 2009).

Structural nested models (SNMs; Robins, 1994) have been proposed to overcome the challenges for causal inference with time-varying confounding. We focus on a class of SNMs for continuous outcomes, namely, structural nested mean models (SNMMs). Most existing works on SNMMs assume discrete-time data-generating processes and require all subjects to be followed at the same prefixed time points, such as months. The literature of discrete-time SNMMs is fruitful; see, for example, Robins (1994), Lok and DeGruttola (2012), and Yang and Lok (2016); 2018). However, observational data are often collected by user-initiated visits to clinics, hospitals, and pharmacies, and data are more likely to be measured at irregularly spaced time points, which are not necessarily the same for all subjects. Such data sources are now commonplace, such as electronic health records, claims databases, disease data registries, and so on. The existing causal framework does not directly apply in such situations, requiring some (possibly arbitrary) discretization of the timeline (Neugebauer *et al.*, 2010). Data preprocessing is quite standard and routine to practitioners, but leads to many unresolved problems: the treatment process depends transparently on the discretization, and therefore the interpretation of SNMMs depends on the definition of time interval (Robins, 1998). Moreover, after discretization, the data may need to be recreated at certain time points. Consider monthly data for example. If a subject had multiple visits within the same month, a common strategy is to take the average of the multiple measures as the observation for a given variable at that month. If a subject had no visit for a given month, one may need to impute the missing observation. Because of such distortions, the resulting data may not satisfy the standard causal consistency or no unmeasured confounding (NUC) assumptions. Consequently, model parameters may not have a causal interpretation.

Our interest is motivated by the observational AIEDRP (Acute Infection and Early Disease Research Program) study (Hecht *et al.*, 2006). This study included a cohort of HIV-infected patients diagnosed during the acute or early stage of disease. Patients initiated highly active antiretroviral therapy (HAART) at various times of the follow-up. Among all AIEDRP patients, 64% of patients initiated the treatment before year 2 with the observed time to treatment initiation ranging continuously from 12 to 282 days. The hypothesis was that deferring therapy may have an increased risk of permanent immune system damage but also a decreased risk of developing drug resistance. Thus, the interest was to estimate the effect of time to initiate HAART on disease progression. However, the complex data also present novel challenges for statistical analysis as summarized below. By study protocol, follow-up visits were scheduled at weeks 2, 4, and 12, and then every 12 weeks thereafter, through week 96. During each follow-up visit, various variables can be measured such as CD4 count and viral load, for which lower CD4 count and higher viral load indicate worse immunological function and disease progression. Treatment initiation can then be determined by the discretion of physicians at follow-up visits. This raises a major concern of time-varying confounding that may obscure the causal effect of time to treatment initiation on disease progression; for example, patients with worse disease progression tend to initiate HAART earlier. Moreover, although the study protocol set visit times in advance, these fixed visit times were not adhered to perfectly in practice. Figure 1 shows the visit patterns from five random patients from the AIEDRP study. Importantly, both the number and the timings of visits differ from one patient to the next. Finally, 45% of patients dropped out of the study before year 2, resulting in right-censored data. It is arguable that censoring due to dropout may depend on the patient's status, which necessitates proper adjustment of censoring.

With irregularly spaced observations, it is more reasonable to assume that the data are generated from continuous-time processes. The work for causal models in continuous-time processes is somewhat sparse; exceptions include, for example, Robins (1998), Lok (2008), Zhang *et al.* (2011), Lok (2017), and Yang *et al.* (2020). Extending the existing causal models with discrete-time processes to continuous-time processes is not trivial. An important challenge lies in time-dependent selection bias or confounding; for example, in a health-related study, sicker patients may visit the doctor more frequently and are more likely to initiate the treatment. To overcome this challenge, following Lok (2008), we treat the observed treatment assignment process as a counting process $N_T(t)$ and assume a martingale condition of NUC on $N_T(t)$ to identify the SNMM parameter. Specifically, the NUC assumption entails that the jumping rate of $N_T(t)$ at t does not depend on future potential outcomes, given the past treatment and covariate history up to t. A practical implication is that the covariate set should be rich enough to include all predictors of outcome and treatment so that we can distinguish the treatment effect and the confounding effect. This assumption was also adopted in Zhang *et al.* (2011), Yang *et al.* (2018), and Yang *et al.* (2020). Lok (2017) provided a strategy of constructing unbiased estimating equations exploiting the relationship between the potential outcome and treatment processes, which leads to a large class of estimators. While this strategy provides unbiased estimators, there is no guidance on how to construct an efficient estimator.

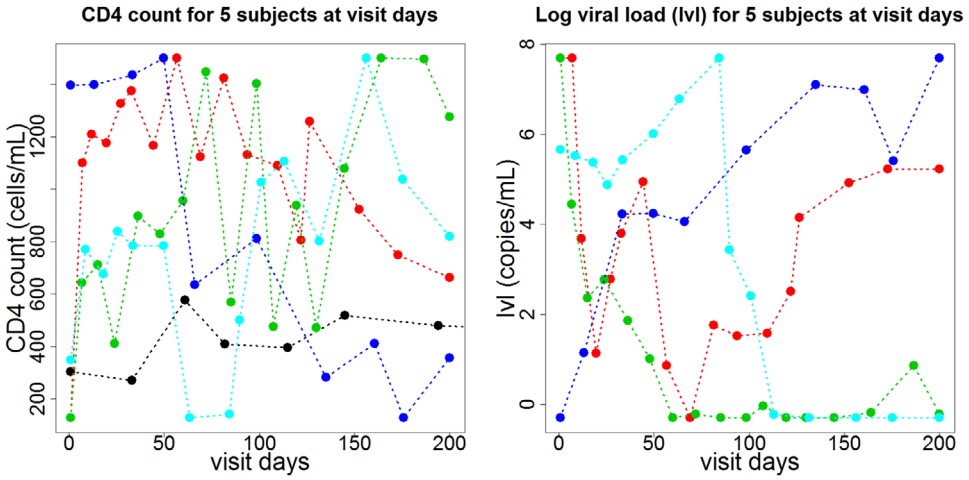

FIGURE 1 CD4 count and log viral load for five random patients measured at irregularly spaced time points, which are colored by patients. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

We establish the new semiparametric efficiency theory for continuous-time SNMMs with irregularly spaced observations. In our problem, the SNMM and NUC assumptions constitute the semiparametric model for the data. Although the NUC assumption does not have any testable implications on the observed-data likelihood, it imposes conditional independence restrictions on the counterfactual outcomes and treatment processes, given the past history, and hence restrictions for the SNMM parameter. To circumvent this complication, we use the variable transformation technique and translate the restrictions into the new variables, which leads to the unconstrained observed-data likelihood. This step allows us to characterize the semiparametric efficiency score (SES) for the SNMM parameter and construct locally efficient estimators that achieve the semiparametric efficiency bound. The estimator requires two nuisance models:

(a) the model of *the potential outcome mean function* conditional on time-varying covariates, and
(b) the model of *the treatment process* conditional on the history of the treatment and covariates.

The proposed estimator of the SNMM parameter is doubly robust in that it is consistent and asymptotically Normal if at least one of the models for the potential outcome mean function and the treatment process is correctly specified. In the AIEDRP database, a large portion of patients dropped out of the study before year 2. Under an ignorable censoring mechanism given the observed history, we show that the complete-case (CC) estimator, that is, the locally efficient estimator applied to the uncensored subjects, remains doubly robust and is optimal among a class of weighting estimators including the inverse probability of censoring weighting (IPCW) estimator (Rotnitzky *et al.*, 2007).

2 | SNMMs IN DISCRETE-TIME PROCESSES

2.1 | Setup, models, and assumptions

We first describe the SNMM in discrete-time processes. We assume that n subjects are followed at prefixed discrete times $t_0 < \cdots < t_{K+1}$ with $t_0 = 0$ and $t_{K+1} = \tau$. We assume that the subjects are simple random samples from a larger population. For simplicity, we suppress the subscript i for subjects. Let L_m be a vector of covariates at time t_m. Let A_m be the treatment indicator at t_m; that is, $A_m = 1$ if the subject was on treatment at t_m and $A_m = 0$ otherwise. We use the overline notation to denote a variable's history; for example, $\overline{A}_m = (A_0, \ldots, A_m)$. We assume that once treatment is initiated, it is never discontinued, so each treatment regime corresponds to one treatment initiation time. Let T be the time to treatment initiation, and let $T = \infty$ if the subject never initiated the treatment during the follow-up. Let Γ be the indicator that the treatment initiation time is less than τ; that is, $\Gamma = 1$ if the subject initiated the treatment before τ and $\Gamma = 0$ otherwise. Let $Y^{(m)}$ be the potential outcome at the end of the study τ, had the subject initiated the treatment at t_m, and let $Y^{(\infty)}$ be the potential outcome at τ had the subject never initiated the treatment during the study follow-up. Let $V_m = (A_{m-1}, L_m)$ be the vector of treatment and covariate for $0 \le m \le K$, where A_{-1} is defined as null. Let Y be the continuous outcome measured at τ. Finally, the subject's full record is $F = (\overline{A}_K, \overline{L}_K, Y)$.

Following Robins (1994) and Lok and DeGruttola (2012), we describe the discrete-time SNMM for the treatment effect as follows.

Assumption 1 (Discrete-time SNMM)**.** For $m \geq 0$, the discrete-time SNMM for the effect of the treatment initiation time is

$$\begin{aligned}\gamma_m(\overline{L}_m) &= \mathbb{E}\left\{ Y^{(m)} - Y^{(\infty)} \mid \overline{A}_{m-1} = \overline{0}, \overline{L}_m \right\} \\ &= \gamma_m(\overline{L}_m; \psi^*); \end{aligned} \tag{1}$$

that is, $\gamma_m(\overline{L}_m; \psi)$ with $\psi \in \mathcal{R}^p$ is a correctly specified model for $\gamma_m(\overline{L}_m)$ with the true parameter value ψ^*.

The general SNMMs in Robins (1994) specify the treatment effects for $\overline{a}_K$ with general patterns. In particular, Robins (1994) focused on modeling $\mathbb{E}\{Y_m^{(\overline{a}_m,\overline{0})} - Y_m^{(\overline{a}_{m-1},\overline{0})} \mid \overline{A}_{m-1} = \overline{a}_{m-1}, \overline{L}_m\}$, where $(\overline{a}_m, \overline{0})$ is the treatment regime of following $\overline{a}_m$ from t_0 to t_m and 0 onwards. This class of models describes one episode of treatment a_m on the shift of the outcome means at t_m given subject's observed treatment and covariates history $(\overline{A}_{m-1} = \overline{a}_{m-1}, \overline{L}_m)$. Model (1) is another class of SNMMs that specify the effects of treatment invitation times (Lok and DeGruttola, 2012). This model characterizes the conditional expectation of the treatment contrasts $Y^{(m)} - Y^{(\infty)}$, given subject's observed history $(\overline{A}_{m-1} = \overline{0}, \overline{L}_m)$. Intuitively, it states that the conditional mean of the outcome is shifted by $\gamma_m(\overline{L}_m; \psi^*)$ had the subject initiated the treatment at t_m comparing to never starting. Therefore, the parameter ψ^* has a causal interpretation. To help understand the model, consider $\gamma_m(\overline{L}_m; \psi^*) = (\psi_1^* + \psi_2^* t_m)(\tau - t_m)^+$, where $\psi^* = (\psi_1^*, \psi_2^*)$ and $c^+ = \max(c, 0)$ for a real number c. This model entails that on average, the treatment would increase the mean of the outcome at τ had the subject initiated the treatment at t_m by $(\psi_1^* + \psi_2^* t_m)(\tau - t_m)^+$, and the magnitude of the increase depends on the duration of the treatment and the treatment initiation time. If $\psi_1^* + \psi_2^* t_m > 0$ and $\psi_2^* < 0$, it indicates the treatment is beneficial and earlier initiation is better. Although we use the same notation ψ^* for the SNMM parameter as Robins (1994), it is important to keep in mind that interpretation of model parameters is tied to the class of SNMMs: Robins (1994) defines the effect of "blipping off" treatment at a single time point, whereas Model (1) defines the effect of removing treatment across all time points.

The following consistency assumption links the observed data to the potential outcomes.

Assumption 2 (Consistency)**.** The observed outcome is equal to the potential outcome under the actual treatment received; that is, $Y = Y^{(T)}$.

If all potential outcomes were observed for each subject, we can directly compare these outcomes to infer the treatment effect; however, the fundamental problem in causal inference is that we cannot observe all potential outcomes for a specific subject (Holland, 1986). In particular, we can observe $Y^{(\infty)}$ only for the subjects who did not initiate the treatment during the follow-up. To overcome this issue, define

$$H(\psi^*) = Y - \gamma_T(\overline{L}_T; \psi^*). \tag{2}$$

Intuitively, $H(\psi^*)$ subtracts the treatment effect $\gamma_T(\overline{L}_T; \psi^*)$ from the observed outcome Y, so it mimics the potential outcome $Y^{(\infty)}$ had the treatment never been initiated. We provide the formal statement as proved in Lok and DeGruttola (2012).

Proposition 1 (Mimicking $Y^{(\infty)}$)**.** *Under Assumption 2, $H(\psi^*)$ mimics $Y^{(\infty)}$, in the sense that $\mathbb{E}\{H(\psi^*) \mid \overline{A}_{m-1} = \overline{0}, A_m, \overline{L}_m\} = \mathbb{E}\{Y^{(\infty)} \mid \overline{A}_{m-1} = \overline{0}, A_m, \overline{L}_m\}$ for $0 \leq m \leq K$.*

We cannot fit the SNMM by a regression model pooled over time, because the model involves the unobserved potential outcomes. Parameter identification requires the NUC assumption (Robins *et al.*, 1992).

Assumption 3 (No unmeasured confounding)**.** $A_m \perp Y^{(\infty)} \mid (\overline{A}_{m-1}, \overline{L}_m)$ for $0 \leq m \leq K$.

Assumption 3 holds if $(\overline{A}_{m-1}, \overline{L}_m)$ contains all prognostic factors for $Y^{(\infty)}$ that affect the treatment decision at t_m for $0 \leq m \leq K$. Under this assumption, the observational study can be conceptualized as a sequentially randomized experiment. Proposition 1 implies that under Assumption 3, for $0 \leq m \leq K$,

$$\begin{aligned}&\mathbb{E}\left\{ H(\psi^*) \mid \overline{A}_{m-1} = \overline{0}, A_m, \overline{L}_m \right\} \\ &= \mathbb{E}\left\{ H(\psi^*) \mid \overline{A}_{m-1} = \overline{0}, \overline{L}_m \right\}; \end{aligned} \tag{3}$$

see, for example, Robins *et al.* (1992). Equation (3) also poses restrictions for ψ^*.

A stronger version of Assumption 3 is $A_m \perp Y^{(k)} \mid (\overline{A}_{m-1}, \overline{L}_m)$ for $0 \leq m \leq K$ and $m \leq k$, requiring the independence between the treatment assignment at t_m and all potential outcomes $Y^{(k)}$ for $k \geq m$, given $(\overline{A}_{m-1}, \overline{L}_m)$. Under this assumption, for $k \leq T$, one may construct the mimicking potential outcome $H(\psi^*) + \gamma^{(k)}(\overline{L}_k; \psi^*)$ for $Y^{(k)}$. The induced restriction for ψ^* from $\mathbb{E}\{Y^{(k)} \mid \overline{A}_{m-1} = \overline{0}, A_m, \overline{L}_m\} = \mathbb{E}\{Y^{(k)} \mid \overline{A}_{m-1} = \overline{0}, \overline{L}_m\}$ is the same as (3). Therefore, it is not necessary to make the stronger version of Assumption 3 in our context.

2.2 | SES for discrete-time SNMMs

The semiparametric model is characterized by the discrete-time SNMM (1) and restriction (3), where the parameter of primary interest is ψ^*. Robins (1994) established the semiparametric efficiency theory for the discrete-time SNMM, following the geometric approach of Bickel *et al.* (1993) by characterizing the nuisance tangent space, its orthogonal complementary space, where all influence functions of regular asymptotically linear (RAL) estimators belong to, and lastly the SES for the SNMM parameter.

Proposition 2 characterizes all influence functions of RAL estimators for ψ^*.

Proposition 2. *For the semiparametric model characterized by the discrete-time SNMM (1) and restriction (3), the influence function space for ψ^* is*

$$\Lambda^{\perp} = \left\{ G(\psi^*; F, c) \ : \ \text{for all } c(m, \overline{V}_m) \in \mathcal{R}^p \right\}, \quad (4)$$

where $\overline{V}_m = (\overline{A}_{m-1}, \overline{L}_m)$ and $G(\psi; F, c) = \sum_{m=1}^{K} c(m, \overline{V}_m)\{A_m - \text{pr}(A_m = 1 \mid \overline{V}_m)\}[H(\psi) - \mathbb{E}\{H(\psi) \mid \overline{V}_m\}]$ indexed by c. To make the notation accurate, the abbreviation c in $G(\psi; F, c)$ means $c(m, \overline{V}_m)$.

The SES, that is, the most efficient one among the class in (4), often does not have a closed-form expression. To fix ideas, consider $\gamma_m(\overline{L}_m; \psi^*) = (\psi_1^* + \psi_2^* t_m)(\tau - t_m)^+$. We now make a working assumption, which extends restriction (3) and allows us to derive an analytical expression of the SES of ψ^*.

Proposition 3 (Discrete-time SES). *Suppose Assumptions 1–3 hold. Suppose further that for $0 \le m \le K$, $\text{var}\{H(\psi^*) \mid \overline{A}_m, \overline{L}_m\} = \text{var}\{H(\psi^*) \mid \overline{V}_m\}$. Then, the SES of ψ^* is*

$$S_{\text{eff}}(\psi^*; F) = G(\psi^*; F, c_{\text{eff}}), \quad (5)$$

where

$$c_{\text{eff}}(m, \overline{V}_m) = \begin{pmatrix} (\tau - t_m) - \mathbb{E}\left\{ \text{dur}(t_m) \mid \overline{A}_m = \overline{0}, \overline{L}_m \right\} \\ t_m(\tau - t_m) - \mathbb{E}\left\{ T \times \text{dur}(t_m) \mid \overline{A}_m = \overline{0}, \overline{L}_m \right\} \end{pmatrix} \left[\text{var}\left\{ H(\psi^*) \mid \overline{V}_m \right\} \right]^{-1},$$

and $\text{dur}(t_m) = \sum_{l=m}^{K-1} A_l(t_{l+1} - t_l)$ is the observed treatment duration from t_m to τ.

3 | SNMMs IN CONTINUOUS-TIME PROCESSES

3.1 | Setup, models, and assumptions

We now extend discrete-time SNMMs in Section 2 to continuous-time SNMMs. We assume that the variables can change their values at any real time between 0 and τ. We assume that all subjects are followed until τ and consider censoring in Section 3.4. Each subject has multiple visit times. Let $N(t)$ be the counting process for the visit times. Let L_t be the multidimensional covariate process. In contrast to the setting with discrete-time data processes, L_t is a vector of covariates at t and additional information of the past visit times up to but not including t. This is because the past visit pattern, for example, the number and frequency of the visit times may be important confounders for the treatment and outcome processes. Let A_t be the binary treatment process. In our motivating application, the treatment can only be initiated at the follow-up visits; that is, if $A_t = 1$, then $N(t) = 1$. We will model the treatment process directly, although one can model first the visit time process and then treatment assignment at the visit times. Define $Y^{(t)}$ as the potential outcome at τ had the subject initiated the treatment at t, and define $Y^{(\infty)}$ as the potential outcome at τ had the subject never initiated the treatment before τ. Let Y be the continuous outcome measured at τ. For regularity, we assume that the processes are Càdlàg processes, that is, the processes are right-continuous with left limits. Let $V_t = (A_{t-}, L_t)$ be the combined treatment and covariate process, where A_{t-} is the treatment information right before t. We use the overline notation to denote a variable's observed history; for example, $\overline{A}_t = \{A_u : 0 \le u \le t, dN(u) = 1\}$. The subject's full record is $F = \{\overline{V}_\tau, (Y^{(t)} : 0 \le t \le \tau)\}$. The observed data for a subject through τ is $D = (\overline{V}_\tau, Y)$.

We assume the continuous-time SNMM as follows.

Assumption 4 (Continuous-time SNMM). For $t \ge 0$, the continuous-time SNMM for the effect of the treatment initiation time is

$$\gamma_t(\overline{L}_t) = \mathbb{E}\left\{ Y^{(t)} - Y^{(\infty)} \mid \overline{L}_t, T \ge t \right\} = \gamma_t(\overline{L}_t; \psi^*); \quad (6)$$

that is, $\gamma_t(\overline{L}_t; \psi)$ with $\psi \in \mathcal{R}^p$ is a correctly specified model for $\gamma_t(\overline{L}_t)$ with the true parameter value ψ^*. Moreover, $Y^{(t)} \sim Y^{(\infty)} + \gamma_t(\overline{L}_t; \psi^*)$ given $(\overline{L}_t, T \ge t)$, where $\sim$ means "is (conditionally) distributed as."

In the continuous-time SNMM (6), ψ^* can be interpreted as the treatment effect rate for the outcome. For the continuous-time SNMM, we assume that given $(\overline{L}_t, T \ge t)$,

the treatment effect only changes the location of the distribution of the outcome but not on other aspects of the distribution such as the variance. This assumption is stronger than the discrete-time SNMM in Assumption 1. But this assumption is weaker than the rank-preserving assumption of $Y^{(t)} = Y^{(\infty)} + \gamma_t(\overline{L}_t; \psi^*)$ considered in Zhang *et al.* (2011). However, the rank preservation may be restrictive in practice, because it implies that for two subjects i and j with the same treatment and covariate history, $Y_i > Y_j$ must imply $Y_i^{(\infty)} > Y_j^{(\infty)}$. We relax this restriction by imposing a distributional assumption. To link the observed outcome to the potential outcomes, we assume that $Y = Y^{(T)}$. Define the mimicking outcome for $Y^{(\infty)}$ as $H(\psi^*) = Y - \gamma_T(\overline{L}_T; \psi^*)$. By Assumption 4, $H(\psi^*) \sim Y^{(\infty)}$, given $(\overline{L}_t, T \geq t)$.

The continuous-time SNMM (6) can model the treatment effect flexibly. For example, the two-parameter model $\gamma_t(\overline{L}_t; \psi^*) = (\psi_1^* + \psi_2^* t)(\tau - t)^+$ entails that the treatment effect depends on the treatment initiation time and the duration of the treatment. To allow for treatment effect modifiers, we can specify an elaborated treatment effect model including time-varying covariates, such as viral load in the blood. For example, one can consider $\gamma_t(\overline{L}_t; \psi^*) = (\psi_1^* + \psi_2^* t + \psi_3^* \text{lvl}_t + \psi_4^* \text{CD4}_t)(\tau - t)^+$, where lvl_t and CD4_t are the log viral load and CD4 count at t. We discuss effect modification and model selection in Section 6.

An important issue with data from user-initiated visits and treatment initiation is the potential selection bias and confounding, for example, sicker patients may visit the doctor more frequently and are likely to initiate treatment earlier. To overcome this issue, we impose the NUC assumption on the treatment process.

Assumption 5 (No unmeasured confounding)**.** The hazard of treatment initiation is

$$\begin{aligned}\lambda_T(t \mid F) &= \lim_{h \to 0} h^{-1} P(t \leq T < t + h, \Gamma = 1 \mid \overline{V}_t, Y^{(\infty)}, T \geq t) \\ &= \lim_{h \to 0} h^{-1} P(t \leq T < t + h, \Gamma = 1 \mid \overline{V}_t, T \geq t),\end{aligned} \tag{7}$$

denoted by $\lambda_T(t \mid \overline{V}_t)$. Because treatment is never discontinued once it is initiated, we impose the condition of $\lambda_T(t \mid \overline{V}_t) = 0$ for $t > T$.

Assumption 5 implies that the hazard of treatment initiation at t depends only on the observed treatment and covariate history $\overline{V}_t$ but not on the future observations and potential outcomes. This assumption holds if the set of historical covariates contains all prognostic factors for the outcome that affect the patient's decision of visiting the doctor and initiating treatment. As an example, in the motivating application, time-invariant characteristics such as age at infection, gender, race, and whether ever used injection drugs are important confounders for the treatment and outcome processes. Moreover, time-varying CD4 and viral load are important confounders. Often, poor disease progression necessitates more frequent follow-up visits and earlier treatment initiation.

The treatment process A_t can also be represented in terms of the counting process $N_T(t)$ and the at-risk process $R_T(t)$ of observing treatment initiation. Let $\sigma(V_t)$ be the σ-field generated by V_t, and let $\sigma(\overline{V}_t)$ be the σ-field generated by $\cup_{u \leq t} \sigma(V_u)$. Under the standard regularity conditions for the counting process, $M_T(t) = N_T(t) - \int_0^t \lambda_T(u \mid \overline{V}_u) R_T(u) du$ is a martingale with respect to the filtration $\sigma(\overline{V}_t)$. Assumption 5 entails that the jumping rate of $N_T(t)$ at t does not depend on $Y^{(\infty)}$, given $\overline{V}_t$. Because $H(\psi^*)$ mimics $Y^{(\infty)}$ in the sense that it has the same distribution as $Y^{(\infty)}$ given $\overline{V}_t$, Assumption 5 also implies that the jumping rate of $N_T(t)$ at t does not depend on $H(\psi^*)$, given $\overline{V}_t$. To be formal, we show in the online supporting information that

$$\lambda_T\{t \mid \overline{V}_t, H(\psi^*)\} = \lambda_T(t \mid \overline{V}_t). \tag{8}$$

Therefore, under the standard regularity conditions, $M_T(t)$ is a martingale with respect to the filtration $\sigma\{\overline{V}_t, H(\psi^*)\}$.

3.2 | SES for continuous-time SNMMs

To estimate the causal parameter precisely, we establish the new semiparametric efficiency theory for the continuous-time SNMMs in parallel to that for the discrete-time SNMMs. We defer all proofs to the online supporting information.

Theorem 1. *For the semiparametric model characterized by the continuous-time SNMM (6) and Assumption 5, the influence function space for ψ^* is $\Lambda^\perp = \{G(\psi^*; F, c)$: for all $c(u, \overline{V}_u) \in \mathcal{R}^p\}$, where*

$$\begin{aligned}G(\psi; F, c) = \int_0^\tau c(u, \overline{V}_u) \Big[H(\psi) - \mathbb{E}\Big\{ H(\psi) \mid \overline{V}_u, T \geq u \Big\} \Big] \\ \times R_T(u) dM_T(u).\end{aligned} \tag{9}$$

From Theorem 1, we can construct a wide class of estimating equation for ψ^* based on $G(\psi; F, c)$ by varying the choice of $c(u, \overline{V}_u)$. The existence of a large number of estimators calls for a principled way to choose $c(u, \overline{V}_u)$ that leads to efficient estimators. Toward this end, we derive the SES for ψ^* by $S_{\text{eff}}(\psi^*; F) = \prod\{S(\psi^*; F) \mid \Lambda^\perp\}$, where

$S(\psi^*;F)$ is the score function of ψ^*. This result motivates efficient estimators of ψ^* in the next subsection.

Theorem 2 (Continuous-time SES). *Under the semiparametric model characterized by the continuous-time SNMM (6) and Assumption 5, the SES of ψ^* is*

$$S_{\text{eff}}(\psi^*;F) = G(\psi^*;F,c_{\text{eff}}), \tag{10}$$

where $G(\psi;F,c)$ is defined in (9), $\dot{H}_u(\psi) = H(\psi) - \mathbb{E}\{H(\psi) \mid \overline{V}_u, T \geq u\}$, and

$$\begin{aligned} c_{\text{eff}}(u,\overline{V}_u) = [&\mathbb{E}\{\partial \dot{H}_u(\psi^*)/\partial\psi \mid \overline{V}_u, T = u\} \\ &- \mathbb{E}\{\partial \dot{H}_u(\psi^*)/\partial\psi \mid \overline{V}_u, T \geq u\}] \\ &\times [\text{var}\{H(\psi^*) \mid \overline{V}_u, T \geq u\}]^{-1}. \end{aligned} \tag{11}$$

To illustrate the theorem, we provide the explicit expression of the SES using an example.

Example 1. Consider $\gamma_t(\overline{L}_t;\psi) = (\psi_1 + \psi_2 t)(\tau - t)^+$. Suppose Assumption 5 holds. The SES of ψ^* is $S_{\text{eff}}(\psi^*;F) = G(\psi^*;F,c_{\text{eff}})$, where

$$\begin{aligned} c_{\text{eff}}(u,\overline{V}_u) = &\begin{pmatrix} (\tau-u)^+ - \mathbb{E}\{(\tau-T)^+ \mid \overline{V}_u, T \geq u\} \\ u(\tau-u)^+ - \mathbb{E}\{T(\tau-T)^+ \mid \overline{V}_u, T \geq u\} \end{pmatrix} \\ &\times [\text{var}\{H(\psi^*) \mid \overline{V}_u, T \geq u\}]^{-1}. \end{aligned} \tag{12}$$

Remark 1. The proposed continuous-time SES contains the discrete-time SES as a special case. If the processes take observations at discrete times $\{t_0,\dots,t_K\}$, then (i) the conditioning event $(\overline{V}_u, T \geq u)$ at t_m is the same as $(\overline{A}_m = \overline{0}, \overline{L}_m)$, (ii) $M_T(t) = N_T(t) - \int_0^t \lambda_T(u \mid \overline{V}_u)R_T(u)du$ at $t = t_m$ becomes $A_m - \text{pr}(A_m = 1 \mid \overline{A}_{m-1} = \overline{0}, \overline{L}_m)$, and $\mathbb{E}\{\partial \dot{H}_t(\psi^*)/\partial\psi \mid \overline{V}_t, T = t\}$ at $t = t_m$ becomes

$$\begin{aligned} &\mathbb{E}\{\partial \dot{H}_m(\psi^*)/\partial\psi \mid \overline{V}_m, T = t_m\} \\ &= -\begin{pmatrix} (\tau-t_m)^+ - \mathbb{E}\left\{\text{dur}(t_m) \mid \overline{A}_m = \overline{0}, \overline{L}_m\right\} \\ t_m(\tau-t_m)^+ - \mathbb{E}\left\{T \times \text{dur}(t_m) \mid \overline{A}_m = \overline{0}, \overline{L}_m\right\} \end{pmatrix}. \end{aligned}$$

Therefore, the continuous-time SES (10) reduces to the discrete-time SES (5).

3.3 | Doubly robust and locally efficient estimators

We first construct a general class of estimators based on the estimating function $G(\psi^*;F,c)$. Because $\mathbb{E}\{G(\psi^*;F,c)\} = 0$, we obtain the estimator of ψ^* by solving

$$\mathbb{P}_n\{G(\psi;F,c)\} = 0. \tag{13}$$

In particular, Equation (13) with c_{eff} provides the semiparametric efficient estimator of ψ^*.

In (13), we assume that the models for the potential outcome mean function $\mathbb{E}\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$ and the treatment process are known. In practice, they are often unknown and must be modeled and estimated from the data. We posit a working model $\mathbb{E}\{H(\psi^*) \mid \overline{V}_u, T \geq u;\beta\}$, such as a linear regression model, where β is a vector of unknown parameters. We also posit a proportional hazards model with time-dependent covariates for the treatment process; that is, $\lambda_T(t \mid \overline{V}_t;\alpha) = \lambda_{T,0}(t)\exp\{\alpha^{\mathrm{T}} W_T(t,\overline{V}_t)\}$, where $\lambda_{T,0}(t)$ is an unknown baseline hazard function, $W_T(t,\overline{V}_t)$ is a prespecified function of t and $\overline{V}_t$, and α is a vector of unknown parameters. Under Assumption 5, we can estimate α and $\lambda_{T,0}(t)$ from the standard software such as "coxph" in R (R Development Core Team, 2012). Fitting the time-dependent proportional hazards model to the data $\{(\overline{V}_{T_i,i}, T_i, \Gamma_i) : i = 1,\dots,n\}$, where $\Gamma_i = I(T_i \leq \tau)$, treating the treatment initiation as the failure event, we obtain the estimators $\widehat{\alpha}$ and $\widehat{\lambda}_{T,0}(t)$. Then, we obtain $\widehat{\lambda}_T(u \mid \overline{V}_u) = \exp\{\widehat{\alpha}^{\mathrm{T}} W_T(u,\overline{V}_u)\} \times \widehat{\lambda}_{T,0}(u)$ and $\widehat{M}_T(t) = N_T(t) - \int_0^t \widehat{\lambda}_T(u \mid \overline{V}_u)R_T(u)du$. As we show below, the resulting estimator $\widehat{\psi}$ achieves the double robustness property.

Theorem 3 (Double robustness). *Suppose the continuous-time SNMM (6) in Assumption 4, and Assumption 5 hold. The estimator $\widehat{\psi}$ solving the estimating equation (13) based on the class of $G(\psi;F,c)$ in (9) by varying $c(u,\overline{V}_u)$ is doubly robust in that it is consistent if at least one of the models for the potential outcome mean function $E\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$ and the treatment process is correctly specified.*

The choice of c does not affect the double robustness but the efficiency of the resulting estimator. For efficiency consideration, we consider c_{eff} in (11). The resulting estimator solving the estimating equation (13) with c_{eff} is locally efficient, in the sense that it achieves the semiparametric efficiency bound if the working models for the treatment process and the potential outcome mean function $E\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$ are correctly specified. Because c_{eff} depends on the unknown distribution, we require additional models for $\mathbb{E}\{(\tau-T)^+ \mid \overline{V}_u, T \geq u\}$ and $\mathbb{E}\{T(\tau-T)^+ \mid \overline{V}_u, T \geq u\}$ to approximate c_{eff}. For example, we can approximate $\mathbb{E}\{(\tau-T)^+ \mid \overline{V}_u, T \geq u\}$ by $P(T \leq \tau \mid \overline{V}_u, T \geq u) \times \mathbb{E}\{\tau - T \mid \overline{V}_u, u \leq T \leq \tau\}$ and each approximated by (logistic) linear models. For $\text{var}\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$, we consider the following options: (i) assume $\text{var}\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$ to be a constant, and (ii) approximate $\text{var}\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$

by the sample variance of $H(\widehat{\psi}_p)$ among subjects with $T \geq u$, where $\widehat{\psi}_p$ is a preliminary estimator. We compare the two options via simulation. Although option (ii) provides a slight efficiency gain in the estimation, for ease of implementation we recommend option (i). Option (i) is common in the generalized estimating equation framework. From here on, we use this option for c and suppress the dependence on c for estimating functions.

Remark 2. Double robustness has appeared for estimators in other contexts of causal inference, such as the augmented inverse probability weighting estimator (AIPW) of the average treatment effect (e.g., Lunceford and Davidian, 2004; Bang and Robins, 2005). Specifically, the AIPW estimator is consistent if either the potential outcome mean function or the propensity score is correctly specified, similar to the requirement for the nuisance functions in Theorem 3. However, the double robustness result in Theorem 3 requires the SNMM to be correctly specified at the outset. Thus, our result requires an additional modeling assumption on the SNMM compared to the typical result for the AIPW estimator.

3.4 | Censoring

As in the AIEDRP study, in most longitudinal observational studies, subjects may drop out of the study prematurely before the end of the study, which renders the data censored at the time of dropout. If the censoring mechanism depends on time-varying prognostic factors, for example, sicker patients drop out of the study with a higher probability than healthier patients, the patients remaining in the study is a biased sample of the full population. We now introduce C to be the time to censoring. Let $X = \min(C, \tau)$ be time to censoring or the end of the study, whichever came first. Let $\delta_C = I(C \geq \tau)$ be the indicator of not censoring before τ. The observed data are $D = (X, \overline{V}_X, \delta_C, \delta_C Y)$.

In the presence of censoring, the estimating equation (13) is not feasible. We assume an ignorable censoring mechanism as follows.

Assumption 6 (Ignorable censoring)**.** The hazard of censoring is

$$\begin{aligned}\lambda_C(t \mid F) &= \lim_{h\to 0} h^{-1} P(t \leq C < t+h \mid F, C \geq t) \\ &= \lim_{h\to 0} h^{-1} P(t \leq C < t+h \mid \overline{V}_t, C \geq t) \quad (14)\end{aligned}$$

denoted by $\lambda_C(t \mid \overline{V}_t)$.

Assumption 6 states that $\lambda_C(t \mid F)$ depends only on the past treatment and covariate history until t, but not on the future variables and potential outcomes. This assumption holds if the set of historical covariates contains all prognostic factors for the outcome that affect the possibility of loss to follow up at t. Under this assumption, the missing data due to censoring are missing at random (Rubin, 1976). From $\lambda_C(t \mid \overline{V}_t)$, we define $K_C(t \mid \overline{V}_t) = \exp\{-\int_0^t \lambda_C(u \mid \overline{V}_u)du\}$, which is the probability of the subject not being censored before t.

We discuss important implications of Assumption 6 on the mimicking potential outcome and the treatment process. First, Assumptions 4 and 6 yield $H(\psi^*) \sim Y^{(\infty)}$, given $(\overline{L}_t, T \geq t, C \geq t)$. Second, under Assumption 6, the hazard of treatment initiation in (7) is equal to $\lim_{h\to 0} h^{-1} P(t \leq T < t+h, \Gamma = 1 \mid \overline{V}_t, T \geq t, C \geq t)$. Redefining T to be the time to treatment initiation, or censoring, or the end of the study, whichever came first, (7) can be estimated by conditioning on $T \geq t$ with the new definition of T. Therefore, the estimating equation (13) restricted to the uncensored subjects remain unbiased. This leads to the CC estimator $\widehat{\psi}$ solving the following equation:

$$\mathbb{P}_n\{\delta_C G(\psi; F)\} = 0. \quad (15)$$

In fact, one can obtain a class of weighting estimators by solving

$$\mathbb{P}_n\left\{\delta_C g(\overline{V}_\tau) G(\psi; F)\right\} = 0 \quad (16)$$

for any weight function $g(\overline{V}_\tau) \in \mathcal{R}$. In particular, choosing $g(\overline{V}_\tau)$ to be $\{K_C(\tau \mid \overline{V}_\tau)\}^{-1}$ leads to the IPCW estimator (Rotnitzky *et al.*, 2007). We show that the CC estimator $\widehat{\psi}$ is optimal among all estimators solving the estimating equation (16). This is because using varying weights reduces the effective sample size compared to constant weights, a classical result in survey sampling (Kish, 1992). Theorem 4 summarizes the asymptotic properties of the CC estimator.

Theorem 4. *Suppose the continuous-time SNMM (6), Assumptions 5 and 6, and regularity conditions in Assumption S1 hold. The CC estimator $\widehat{\psi}$ solving the estimating equation (15) is doubly robust in that it is consistent and asymptotically Normal if at least one of the models for the potential outcome mean function $E\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$ and the treatment process is correctly specified. Moreover, if both nuisance models are correctly specified, $\widehat{\psi}$ achieves the smallest variance among the class of estimators solving the estimating equation (16).*

The asymptotic Normal distribution is presented in the online supporting information and is agnostic about

whether the potential outcome mean function $E\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$ or the treatment process is correctly estimated. However, it is difficult to use the asymptotic variance formula for variance estimation because it requires approximating additional nuisance functions. From Theorem 4, under the conditions that ensure double robustness, $\widehat{\psi}$ is asymptotically linear with a Normal limiting distribution, and therefore, we can use the nonparametric bootstrap for variance estimation.

4 | SIMULATION STUDY

We now evaluate the finite-sample performance of the proposed estimator on simulated data sets with two objectives. First, we assess the double robustness and efficiency of the proposed estimator based on the SES, compared with some preliminary estimator. Second, to demonstrate the impact of data discretization as commonly done in practice, we include the g-estimator applied to the preprocessed data. We simulate 1000 data sets under two settings with and without censoring with sample size $n = 1000$. Additional simulation results with $n = 2000$ are presented in the online supporting information.

In Setting I, we generate two covariates, one time-independent (L_{TI}) and one time-dependent (L_{TD}). The time-independent covariate L_{TI} is generated from a Bernoulli distribution with mean equal to 0.55. The time-dependent covariate is $L_{TD,t} = l_1 \times I(0 \leq t < 0.5) + l_2 \times I(0.5 \leq t < 1) + l_3 \times I(1 \leq t < 1.5) + l_4 \times I(1.5 \leq t \leq 2)$, where $(l_1, l_2, l_3, l_4)^{\mathrm{T}}$ is a 1×4 row vector generated from a multivariate Normal distribution with mean equal to (0,0,0,0) and covariance equal to $0.7^{|i-j|}$ for $i, j = 1, \ldots, 4$. We assume that the time-dependent variable remains constant between measurements. The maximum follow-up time is $\tau = 2$ (in year). We generate the time to treatment initiation T with the hazard rate $\lambda_T(t \mid \overline{V}_t) = \lambda_{T,0}(t) \exp(\alpha_1 \times L_{TI} + \alpha_2 L_{TD,t})$ with $\lambda_{T,0}(t) = 0.4$, $\alpha_1 = 0.15$, and $\alpha_2 = 0.8$. We generate T according to the time-dependent model sequentially. This is because the hazard of treatment initiation in the time interval from $t_1 = 0$ to $t_2 = 0.5$ differs from the hazard of treatment initiation in the next interval and so on; see the online supporting information for details. We let $Y^{(\infty)} = L_{TD,\tau}$ be the potential outcome had the subject never initiated the treatment before τ. The observed outcome is $Y = Y^{(\infty)} + \gamma_T(\overline{V}_T; \psi^*)$, where $\gamma_t(\overline{V}_t; \psi^*) = (\psi_1^* + \psi_2^* t)(\tau - t)^+$ with $\psi_1^* = 15$ and $\psi_2^* = -1$.

We consider the following estimators with details for the nuisance models and their estimation presented in the online supporting information:

a) A preliminary estimator $\widehat{\psi}_p$ solves (13) with $E\{H(\psi^*) \mid \overline{V}_u, T \geq u\} \equiv 0$ and $c(u, \overline{V}_u) = (1, u)^{\mathrm{T}}(\tau - u)^+ - \mathbb{E}\{(1, T)^{\mathrm{T}}(\tau - T)^+ \mid \overline{V}_u, T \geq u\}$. Therefore, $\widehat{\psi}_p$ corresponds to the proposed estimator with a misspecified model for $\mathbb{E}\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$.
b) The proposed estimator $\widehat{\psi}_{\mathrm{cont},1}$ solves (13) , where we replace $\mathrm{var}\{H(\psi) \mid \overline{V}_u, T \geq u\}$ by a constant.
c) The proposed estimator $\widehat{\psi}_{\mathrm{cont},2}$ solves (13) , where we obtain $\widehat{\mathrm{var}}\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$ by the empirical variance of $H(\widehat{\psi}_p) - \mathbb{E}\{H(\widehat{\psi}_p) \mid \overline{V}_u, T \geq u; \widehat{\beta}\}$, restricted to subjects with $T \geq u$.
d) The g-estimator $\widehat{\psi}_{\mathrm{disc},g}$ in Section 2 applies to the monthly data after discretization with 24 equally spaced time points from 0 to τ. For $m \geq 1$, at the mth time point t_m, L_m is the average of L_t from $t_{m-1} \leq t \leq t_m$, A_m is the indicator of whether the treatment is initiated before t_m, and the time to treatment initiation T is t_m if $A_m = 1$ and $\overline{A}_{m-1} = \overline{0}$. The g-estimator solves the estimating equation based on (5), where the nuisance models are estimated similar to what are used for $\widehat{\psi}_{\mathrm{cont},1}$ but with the reshaped data.

To investigate the double robustness in Theorem 3, we consider two models for estimating M_T: the correctly specified proportional hazards model with both time-independent and time-dependent covariates; and the misspecified proportional hazards model with only time-independent covariate. For all estimators, we use the bootstrap for variance estimation with the bootstrap size 100.

Table 1 shows the simulation results in Setting I. Under Scenario (i), when the model for the treatment process is correctly specified, $\widehat{\psi}_p$, $\widehat{\psi}_{\mathrm{cont},1}$, and $\widehat{\psi}_{\mathrm{cont},2}$ show small biases. As a result, the coverage rates are close to the nominal level. Under Scenario (ii), when the model for the treatment process is misspecified, $\widehat{\psi}_p$ shows large biases, but $\widehat{\psi}_{\mathrm{cont},1}$ and $\widehat{\psi}_{\mathrm{cont},2}$ still show small biases. Moreover, the root mean squared errors of $\widehat{\psi}_{\mathrm{cont},1}$ and $\widehat{\psi}_{\mathrm{cont},2}$ decrease as the sample size increases; see the additional simulation results in the online supporting information. This confirms the double robustness of the proposed estimators. The proposed estimator $\widehat{\psi}_{\mathrm{cont},2}$ with $\widehat{\mathrm{var}}\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$ produces slightly smaller standard errors; however, this reduction is not large. In practice, we recommend $\widehat{\psi}_{\mathrm{cont},1}$ because of its simpler implementation than $\widehat{\psi}_{\mathrm{cont},2}$. We note large biases in the g-estimator, which illustrates the consequence of data preprocessing for the subsequent analysis.

In Setting II, we further generate the time to censoring C with the hazard rate $\lambda_C(t \mid \overline{V}_t) = \lambda_{C,0}(t) \exp(\eta_1 L_{TI} + \eta_2 L_{TD,t})$, with $\lambda_{C,0}(t) = 0.2$, $\eta_1 = 0.15$, and $\eta_2 = 0.35$. In the presence of censoring, we consider the three estimators (a), (b), and (d) considered in Setting I applied to the

TABLE 1 Simulation results in Setting I without censoring based on 1000 simulated data sets: the Monte Carlo bias, standard error, root mean square error of the estimators, and coverage rate of 95% confidence intervals

		Bias ($\times 10^2$)		SE ($\times 10^2$)		rMSE ($\times 10^2$)		CR ($\times 10^2$)	
	Method	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*
Scenario (i): Model for M_T (✓)									
Model for POM (×)	$\widehat{\psi}_p$	0.3	−0.1	5.3	9.6	5.3	9.6	95.0	94.0
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	0.2	0.1	5.0	8.9	5.0	8.9	95.4	94.0
	$\widehat{\psi}_{\text{cont},2}$	0.2	0.1	4.9	8.7	4.9	8.7	95.3	94.4
–	$\widehat{\psi}_{\text{disc,g}}$	28.6	34.5	6.0	10.5	29.3	36.1	0.0	7.2
Scenario (ii): Model for M_T (×)									
Model for POM (×)	$\widehat{\psi}_p$	7.4	20.2	5.2	9.9	9.1	22.5	68.8	44.6
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	0.5	0.5	5.1	9.1	5.1	9.1	95.4	94.0
	$\widehat{\psi}_{\text{cont},2}$	0.5	0.4	5.1	9.0	5.1	9.0	95.0	95.4
–	$\widehat{\psi}_{\text{disc,g}}$	27.7	38.6	5.9	10.2	28.4	40.0	0.2	3.4

Note "POM" means the potential outcome mean function $E\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$; ✓ (is correctly specified), and × (is misspecified).

uncensored subjects with weighting; that is, the estimators solving the corresponding estimating equations (16) weighted by $g(\overline{V}_\tau)$. To investigate the robustness and optimality of $g(\overline{V}_\tau) = 1$ in Theorem 4, we consider $g(\overline{V}_\tau) = 1$ and $g(\overline{V}_\tau) = \{\widehat{K}_C(\tau \mid \overline{V}_\tau)\}^{-1}$ with two models for estimating K_C: the correctly specified proportional hazards model with $(L_{TI}, L_{TD,t})$, and the misspecified proportional hazards model with $(L_{TI}, L_{TD,t}^2)$.

Table 2 shows the simulation results in Setting II. Under scenarios when either the model for the potential outcome mean function $\mathbb{E}\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$ or the model for the treatment process is correctly specified, the estimators show small biases, regardless of the specification of $g(\overline{V}_\tau)$. Moreover, if both models for the potential outcome mean function and the treatment process are misspecified, $\widehat{\psi}_p$ shows large biases. Under the same model specification, the CC estimator with $g(\overline{V}_\tau) = 1$ is more efficient than the IPCW estimator with a correctly specified censoring model which is more efficient than the IPCW estimator with a misspecified censoring model. This confirms the optimality of $g(\overline{V}_\tau) = 1$ in Theorem 4. Again, the discretized g-estimator shows large biases across all scenarios.

5 | ESTIMATING THE EFFECT OF TIME TO INITIATING HAART

We apply our method to the observational AIEDRP database consisting of 1762 HIV-positive patients diagnosed during acute and early infection (Hecht *et al.*, 2006). This data set was previously used by Lok and DeGruttola (2012) and Yang and Lok (2016; 2018); all these methods were based on the monthly data after discretization. As discussed in the introduction, the observations from the original data are collected by user-initiated visits and are irregularly spaced (Hecht *et al.*, 2006). Figure 1 shows the visit times for five random patients. As can be seen, we have irregular visits, and the number and frequency of visits vary from patient to patient. We aim to estimate the average causal effect of the time to HAART initiation on the mean CD4 count at year 2 after HIV infection directly on the basis of the original data without discretization . The outcome variable Y is the CD4 count measured by the end of year 2, with the interquantile range from 443 to 794 cells/mm^3. The observed time to treatment initiation ranges continuously from 12 to 282 days. To ensure the NUC and ignorable censoring assumptions hold, we include the following baseline and time-varying covariates: age at infection, gender, race, injection drug ever/never, and measured CD4 count and log viral load at follow-up visits. We assume a continuous-time SNMM $\gamma(\overline{V}_u; \psi^*) = (\psi_1^* + \psi_2^* t)(\tau - t)^+$, where ψ_2^* quantifies the impact of time to treatment initiation. The rationale for this modeling choice is because the duration of treatment may well be predictive of its effect.

We consider the proposed CC estimators $\widehat{\psi}_{\text{cont},1}$ and $\widehat{\psi}_{\text{cont},2}$ specified in Section 4 applied to the uncensored subjects. The estimation procedure requires specifying and fitting nuisance models, which we discuss below.

Model for the potential outcome mean function. $E\{H(\widehat{\psi}_p) \mid \overline{V}_u,\ T \geq u;\ \beta\}$ is a linear regression model where the covariates include age, male, race, injdrug, CD4_u, lvl_u, $\text{CD4}_u^{3/4}(\tau - u)$, $\text{CD4}_u^{3/4} \times (\tau - u) \times \text{age}$, $\text{CD4}_u^{3/4} \times (\tau - u) \times \text{male}$, $\text{CD4}_u^{3/4} \times (\tau - u) \times \text{race}$, $\text{CD4}_u^{3/4} \times (\tau - u) \times \text{injdrug}$, $\text{CD4}_u^{3/4} \times (\tau - u) \times \text{lvl}_u$, CD4slope_u measured, $\text{CD4slope}_u \times (\tau - u)^{1/2}$ $(6 - u)^+$, and $(36 - u^2)^+$. This model specification is motivation based on the substantive literature; see, for example, May *et al.* (2009).

TABLE 2 Simulation results in Setting II with censoring based on 1000 simulated data sets: the Monte Carlo bias, standard error, root mean square error of the estimators, and coverage rate of 95% confidence intervals

		Bias ($\times 10^3$)		SE ($\times 10^3$)		rMSE ($\times 10^3$)		CR ($\times 10^2$)	
($n = 1000$)	Method	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*
Scenario (i): Model for M_T (✓) and $g \equiv 1$									
Model for POM (×)	$\widehat{\psi}_p$	0.4	−1.4	65.4	110.9	65.4	110.9	95.7	94.6
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	−0.1	3.4	65.0	106.5	65.0	106.6	95.0	94.5
–	$\widehat{\psi}_{\text{disc},g}$	276.2	315.6	74.5	121.5	286.0	338.2	8.7	10.9
Scenario (ii): Model for M_T (×) and $g \equiv 1$									
Model for POM (×)	$\widehat{\psi}_p$	63.7	177.5	64.4	110.6	90.6	209.2	69.1	43.5
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	6.5	5.3	65.0	107.5	65.4	107.7	94.2	94.5
–	$\widehat{\psi}_{\text{disc},g}$	281.2	318.7	74.4	120.9	290.9	340.9	7.4	8.8
Scenario (iii): Model for M_T (✓) and $g = K_C^{-1}$(✓)									
Model for POM (×)	$\widehat{\psi}_p$	−0.3	−0.3	66.8	114.6	66.8	114.6	95.7	95.5
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	−0.6	5.2	65.7	109.4	65.7	109.5	95.2	94.2
–	$\widehat{\psi}_{\text{disc},g}$	273.4	312.4	75.0	124.1	283.5	336.2	10.3	12.0
Scenario (iv): Model for M_T (×) and $g = K_C^{-1}$(✓)									
Model for POM (×)	$\widehat{\psi}_p$	63.5	160.0	63.2	109.8	89.6	194.0	63.1	48.3
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	10.2	2.5	66.2	108.8	67.0	108.9	94.1	94.7
–	$\widehat{\psi}_{\text{disc},g}$	283.7	311.4	76.0	124.7	293.7	335.4	11.8	13.2
Scenario (v): Model for M_T (✓) and $g = K_C^{-1}$(×)									
Model for POM (×)	$\widehat{\psi}_p$	0.2	−1.5	67.3	114.0	67.3	114.0	95.2	95.6
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	−0.3	3.7	66.7	109.3	66.7	109.4	94.6	94.3
–	$\widehat{\psi}_{\text{disc},g}$	273.4	312.4	75.0	124.1	283.5	336.2	8.4	10.2
Scenario (vi): Model for M_T (×) and $g = K_C^{-1}$(×)									
Model for POM (×)	$\widehat{\psi}_p$	72.6	191.8	66.1	113.6	98.2	223.0	61.4	32.2
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	13.7	4.1	66.9	107.6	68.3	107.7	94.2	94.6
–	$\widehat{\psi}_{\text{disc},g}$	283.7	311.4	76.0	124.7	293.7	335.4	7.7	9.3

Note "POM" means the potential outcome mean function $E\{H(\psi^*) \mid \overline{V}_u, T \geq u\}$; ✓ (is correctly specified), and × (is misspecified).

Model for the treatment process. The model for the treatment process (M_T) is a time-dependent proportional hazards model adjusting for gender, age (age at infection), race (white non-Hispanic race), injdrug (injection drug ever/never), $\text{CD4}_u^{1/2}$ (square root of current CD4 count), lvl_u (log viral load), days from last visit_u (number of days since the last visit), first visit_u (whether the visit is the first visit), and second visit_u (whether the visit is the second visit).

Other nuisance models. $E(\tau - T \mid \overline{L}_u, T \geq u)$ and $E\{T(\tau - T) \mid \overline{L}_u, T \geq u)\}$ are linear regression models where the covariates include u, $(\tau - u)$, male$\times(\tau - u)$, age$\times(\tau - u)$, race$\times(\tau - u)$, injdrug$\times(\tau - u)$, $\text{CD4}_u^{1/2}\times(\tau - u)$, $\text{lvl}_u\times(\tau - u)$, days from last $\text{visit}_u \times (\tau - u)$, first $\text{visit}_u \times (\tau - u)$, and second $\text{visit}_u \times (\tau - u)$.

We use bootstrap for variance estimation with the bootstrap size 100 and compute the 95% Wald confidence interval. We also include the discretized g-estimator (Lok and DeGruttola, 2012) applied to the monthly data. Table 3 shows the results for the effect of time to HAART initiation on the CD4 count at year 2. We note only slight differences in the point estimates between the proposed estimators. The discretized g-estimator is much larger than the proposed estimators for ψ_1^*, but all estimators have similar

TABLE 3 Results of the effect of time to HAART initiation on the CD4 count at year 2

Method	Est	SE	Lower 0.95	Upper 0.95	p-Val
	ψ_1^* cells/mm^3 per month				
Proposed 1: $\widehat{\psi}_{\text{cont},1}$	14.2	1.0	12.2	16.1	0.000
Proposed 2: $\widehat{\psi}_{\text{cont},2}$	14.4	1.0	12.4	16.3	0.000
Disc g-formula: $\widehat{\psi}_{\text{disc},g}$	24.9	1.2	22	28	0.000
	ψ_2^* cells/mm^3 per month^2				
Proposed 1: $\widehat{\psi}_{\text{cont},1}$	−0.96	0.28	−1.51	−0.41	0.000
Proposed 2: $\widehat{\psi}_{\text{cont},2}$	−0.97	0.27	−1.49	−0.45	0.000
Disc g-formula: $\widehat{\psi}_{\text{disc},g}$	−0.73	0.65	−1.50	0.10	0.080

results for ψ_2^*. The results show that earlier HAART initiation is better in increasing CD4 counts. Our estimators respect the underlying continuous-time nature of data processes. Based on our results, on average, initiation of HAART at the time of infection ($t = 0$) can increase CD4 counts at year 2 by 14.2 cells/mm^3 per month $\times$ 24 months $\approx$ 341 cells/mm^3; while initiation of HAART 3 months after the time of infection can increase CD4 counts at year 2 by $(14.2 - 0.96 \times 3) \times (24 - 3) \approx 238$ cells/mm^3. In the supporting information, we conduct a sensitivity analysis using an elaborated SNMM with possible treatment effect modifiers. The analysis also shows earlier HAART initiation is better in increasing CD4 counts, although the result becomes not significant possibly due to the increased number of the SNMM parameter. Finally, we add a caveat that we require the SNMMs to be correctly specified and a formal goodness-of-fit test for model assessment will be our future work.

6 | DISCUSSION

In this paper, we have developed a new semiparametric estimation framework for continuous-time SNMMs to evaluate treatment effects with irregularly spaced longitudinal observations under the assumptions of NUC and ignorability of censoring. We do not require specifying the full distribution of the covariate, treatment, outcome, and censoring processes. Our approach achieves a double robustness property requiring the correct specification of either the model for the potential outcome mean function or the model for the treatment process, regardless of whether or not the model for the censoring process is correctly specified. As discussed previously, the key assumptions hold if all variables are measured that are related to both treatment and outcome and that are related to both censoring and outcome. Although essential, they are not verifiable based on the observed data but rely on subject matter experts to assess their plausibility.

The proposed framework is also applicable to the analysis of patient-reported outcomes in pragmatic clinical trials. Although trial protocols often require collecting outcomes at prefixed time points after randomization, patients may delay or even miss their visits in practice, resulting in irregular-spaced observations. One important implication of the proposed framework for future trial designs is to collect a sufficiently rich set of variables that predict the treatment and censoring processes and ensure the required assumptions hold.

There are several directions for future work: (i) we will extend the continuous-time SNMMs framework to other types of outcomes such as binary outcomes and survival outcomes; and (ii) we will develop a variable selection procedure for identifying effect modifiers. The insight is that we have a larger number of estimating functions than the number of parameters. The problem for effect modifiers selection falls into the recent work of Chang *et al.* (2018) on high-dimensional statistical inferences with overidentification; (iii) a goodness-of-fit test using overidentification (Yang and Lok, 2016). can also be developed to assess the assumption for the SNMM; and (iv) we will develop sensitivity analyses to assess the impact of possible uncontrolled confounding (Yang and Lok, 2018).

ACKNOWLEDGMENTS

The author would like to thank Anastasios A. Tsiatis for insightful and fruitful discussions. The author is grateful to the patients who volunteered for AIEDRP and the AIEDRP study team. Yang is partially supported by the NSF grant DMS 1811245, NCI grant P01 CA142538, NIA grant 1R01AG066883, and NIEHS grant 1R01ES031651.

DATA AVAILABILITY STATEMENT

The data that support the findings of this article are available from the Acute Infection and Early Disease Research Program (AIEDRP) study team. Data are available from the author with the permission of the AIEDRP study team.

ORCID

Shu Yang https://orcid.org/0000-0001-7703-707X

REFERENCES

Bang, H. and Robins, J.M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.

Bickel, P.J., Klaassen, C., Ritov, Y. and Wellner, J. (1993) *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Chang, J., Tang, C.Y. and Wu, T.T. (2018) A new scope of penalized empirical likelihood with high-dimensional estimating equations. *Annals of Statistics*, 46, 3185–3216.

Hecht, F.M., Wang, L., Collier, A., Little, S., Markowitz, M., Margolick, J. et al. (2006) A multicenter observational study of the potential benefits of initiating combination antiretroviral therapy during acute HIV infection. *Journal of Infectious Diseases*, 194, 725–733.

Holland, P.W. (1986) Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960.

Kish, L. (1992) Weighting for unequal pi. *Journal of Official Statistics*, 8, 183–200.

Lok, J.J. (2017) Mimicking counterfactual outcomes to estimate causal effects. *Annals of Statistics*, 45, 461–499.

Lok, J.J. (2008) Statistical modeling of causal effects in continuous time. *Annals of Statistics*, 36, 1464–1507.

Lok, J.J. and DeGruttola, V. (2012) Impact of time to start treatment following infection with application to initiating HAART in HIV-positive patients. *Biometrics*, 68, 745–754.

Lunceford, J.K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23, 2937–2960.

May, M., Wood, R., Myer, L., Taffé, P., Rauch, A., Battegay, M. et al. (2009) CD4 T-cell declines by ethnicity in untreated HIV-1 infected patients in South Africa and Switzerland. *Journal of Infectious Diseases*, 200, 1729–1735.

Neugebauer, R., Silverberg, M.J. and van der Laan, M.J. (2010) *Observational study and individualized antiretroviral therapy initiation rules for reducing cancer incidence in HIV-infected patients*. Technical Report, U.C. Berkeley Division of Biostatistics Working Paper Series: Working Paper 272.

R Development Core Team. (2012) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Robins, J.M. (1994) Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics – Theory and Methods*, 23, 2379–2412.

Robins, J.M. (1998) Correction for non-compliance in equivalence trials. *Statistics in Medicine*, 17, 269–302.

Robins, J.M., Blevins, D., Ritter, G. and Wulfsohn, M. (1992) G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology*, 3, 319–336.

Robins, JM, Hernán, MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice, G, Davidian, M, Verbeke, G, Molenberghs, G, editors. *Advances in Longitudinal Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press; (2009). 553–599.

Rotnitzky, A., Farall, A., Bergesio, A. and Scharfstein, D. (2007) Analysis of failure time data under competing censoring mechanisms. *Journal of Royal Statistical Society Series B*, 69, 307–327.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63, 581–592.

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

Yang, S. and Lok, J.J. (2018) Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statistica Sinica*, 28, 1703–1723.

Yang, S., Pieper, K. and Cools, F. (2020) Semiparametric estimation of structural failure time model in continuous-time processes. *Biometrika*, 107, 123–136.

Yang, S., Tsiatis, A.A. and Blazing, M. (2018) Modeling survival distribution as a function of time to treatment discontinuation: a dynamic treatment regime approach. *Biometrics*, 74, 900–909.

Zhang, M., Joffe, M.M. and Small, D.S. (2011) Causal inference for continuous-time processes when covariates are observed only at discrete times. *Annals of Statistics*, 39, 131–173.

SUPPORTING INFORMATION

Web Appendices and Tables referenced in Sections 3—5 and R code for implementing the proposed method are available with this article at the *Biometrics* website on Wiley Online Library.

How to cite this article: Yang S. Semiparametric estimation of structural nested mean models with irregularly spaced longitudinal observations. *Biometrics*. 2021;1–13.
https://doi.org/10.1111/biom.13471

Supporting information for "Semiparametric estimation of structural nested mean models with irregularly spaced longitudinal observations" by Yang

Supporting information includes proofs, technical details and additional simulation and sensitivity analysis in the application.

Web Appendix 1. Proofs

Web Appendix 1.1 *Proof of (8)*

First, we express

$$\begin{aligned}
& \lambda_T\{t \mid \overline{V}_t, Y^{(\infty)}\} = \lim_{h\to 0} h^{-1} P\{t \leqslant T < t+h \mid \overline{V}_t, Y^{(\infty)}, T \geqslant t\} \\
= & \lim_{h\to 0} h^{-1} \frac{f\{Y^{(\infty)} \mid \overline{V}_t, t \leqslant T < t+h\} P\{t \leqslant T < t+h \mid \overline{V}_t, T \geqslant t\}}{f\{Y^{(\infty)} \mid \overline{V}_t, T \geqslant t\}} \\
= & \lim_{h\to 0} h^{-1} \frac{f\{H(\psi^*) \mid \overline{V}_t, t \leqslant T < t+h\} P\{t \leqslant T < t+h \mid \overline{V}_t, T \geqslant t\}}{f\{H(\psi^*) \mid \overline{V}_t, T \geqslant t\}} \\
= & \lim_{h\to 0} h^{-1} P\{t \leqslant T < t+h \mid \overline{V}_t, H(\psi^*), T \geqslant t\} \\
= & \lambda_T\{t \mid \overline{V}_t, H(\psi^*)\},
\end{aligned}$$

where the second equality follows by the Bayes rule, and the third equality follows by Model (6) which implies that the distribution of $\{\overline{V}_t, Y^{(\infty)}\}$ is the same as the distribution of $\{\overline{V}_t, H(\psi^*)\}$.

Second, by Assumption 5, $\lambda_T\{t \mid \overline{V}_t, Y^{(\infty)}\} = \lambda_T(t \mid \overline{V}_t)$. Therefore, $\lambda_T\{t \mid \overline{V}_t, H(\psi^*)\} = \lambda_T\{t \mid \overline{V}_t, Y^{(\infty)}\} = \lambda_T(t \mid \overline{V}_t)$.

Web Appendix 1.2 *General semiparametric efficiency theory*

We present the general semiparametric efficiency theory. Suppose the data consist of n independent and identically distributed random variables $F_1, \ldots, F_n$. We consider regular asymptotically linear (RAL) estimators $\widehat{\psi}_n$ for ψ^* as

$$n^{1/2}(\widehat{\psi}_n - \psi^*) = n^{1/2}\mathbb{P}_n\Phi(F) + o_{\mathbb{P}}(1), \tag{S1}$$

where $\mathbb{P}_n$ denotes the empirical mean; i.e., $\mathbb{P}_n\Phi(F) = n^{-1}\sum_{i=1}^{n}\Phi(F_i)$, $\Phi(F)$ is called the influence function of $\widehat{\psi}_n$, with mean zero and finite and non-singular variance. Because ψ^* is p-dimensional, $\Phi(F)$ is also p-dimensional. From (S1), the asymptotic variance of $n^{1/2}(\widehat{\psi}_n - \psi^*)$ is equal to the variance of its influence function. As a result, to construct the efficient RAL estimator, it suffices to find the influence function with the smallest variance.

To do this, we describe a geometric approach of Bickel et al. (1993). Consider the Hilbert space $\mathcal{H}$ of all p-dimensional, mean-zero finite variance measurable functions of F, denoted by $h(F)$, equipped with the covariance inner product $< h_1, h_2 >= \mathbb{E}\left\{h_1(F)^{\mathrm{T}}h_2(F)\right\}$ and the norm $||h|| = \mathbb{E}\left\{h(F)^{\mathrm{T}}h(F)\right\}^{1/2} < \infty$. Bickel et al. (1993) stated that influence functions for RAL estimators lie in the orthogonal complement of the nuisance tangent space in $\mathcal{H}$. To motive the concept of the nuisance tangent space for a semiparametric model, we first consider a fully 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 $S(\theta^*;F) = \partial\log f(F;\psi^*,\theta)/\partial\theta$, both evaluated at the true values (ψ^*,θ^*), respectively. For a parametric model, the nuisance tangent space Λ is the linear space in $\mathcal{H}$ spanned by the q-dimensional nuisance score vector $S(\theta;F)$. For semiparametric models, where the nuisance parameter is infinite-dimensional, the nuisance tangent space Λ is defined as the mean squared closure of all parametric sub-model nuisance tangent spaces. The efficient score $S_{\text{eff}}(F)$ for the semiparametric model is the projection of $S(\psi^*;F)$ onto the orthogonal complementary space of the nuisance tangent space $\Lambda^{\perp}$; i.e., $S_{\text{eff}}(F) = \prod\left[S(\psi^*;F) \mid \Lambda^{\perp}\right]$, where $\prod$ is the projection operator in the Hilbert space. The efficient influence function is $\Phi_{\text{eff}}(F) = \left[\mathbb{E}\left\{S_{\text{eff}}(F)S_{\text{eff}}(F)^{\mathrm{T}}\right\}\right]^{-1}S_{\text{eff}}(F)$, with the variance $\left[\mathbb{E}\left\{S_{\text{eff}}(F)S_{\text{eff}}(F)^{\mathrm{T}}\right\}\right]^{-1}$, which achieves the semiparametric efficiency bound (Bickel et al., 1993). From this geometric point of view, to derive efficient semiparametric estimators for ψ^*, it suffices to find the efficient score $S_{\text{eff}}(F)$.

Web Appendix 1.3 *Proof of Theorem 1*

First, we characterize the semiparametric likelihood function of ψ^* based on a single variable $D = (\overline{V}_\tau, Y)$. The semiparametric likelihood is

$$f_D\left(\overline{V}_\tau, Y\right) = \left\{\frac{\mathrm{d}H(\psi^*)}{\mathrm{d}Y}\right\} f_{\{\overline{V}_\tau, H(\psi^*)\}}\{\overline{V}_\tau, H(\psi^*)\} = f_{\{\overline{V}_\tau, H(\psi^*)\}}\{\overline{V}_\tau, H(\psi^*)\}, \tag{S2}$$

where the first equality follows by the transformation of D to $\{\overline{V}_\tau, H(\psi^*)\}$, and the second equality follows because $\mathrm{d}H(\psi^*)/\mathrm{d}Y = 1$. To express (S2) further, we let the observed times to treatment initiation among the n subjects be $v_0 = 0 < v_1 < \cdots < v_M$. By Assumption 5 and (8), we express

$$\begin{aligned}
f_D\left(\overline{V}_\tau, Y; \psi^*, \theta\right) &= f\left\{H(\psi^*); \theta_1\right\} \prod_{k=1}^{M} f\left\{L_{v_k} \mid \overline{A}_{v_{k-1}} = \overline{0}, \overline{L}_{v_{k-1}}, H(\psi^*); \theta_2\right\} \\
&\quad \times \prod_{v=v_1}^{v_M} f\left\{A_{v_k} \mid \overline{A}_{v_{k-1}} = \overline{0}, \overline{L}_{v_k}, H(\psi^*); \theta_3\right\}, \\
&= f\left\{H(\psi^*); \theta_1\right\} \prod_{k=1}^{M} f\left\{L_{v_k} \mid \overline{A}_{v_{k-1}} = \overline{0}, \overline{L}_{v_{k-1}}, H(\psi^*); \theta_2\right\} \\
&\quad \times \prod_{v=v_1}^{v_M} f\left\{A_{v_k} \mid \overline{A}_{v_{k-1}} = \overline{0}, \overline{L}_{v_k}; \theta_3\right\} \\
&= f\left\{H(\psi^*); \theta_1\right\} \prod_{k=1}^{M} f\left\{L_{v_k} \mid \overline{A}_{v_{k-1}} = \overline{0}, \overline{L}_{v_{k-1}}, H(\psi^*); \theta_2\right\} \\
&\quad \times f(T, \Gamma \mid \overline{V}_T; \theta_3),
\end{aligned} \tag{S3}$$

where $\theta = (\theta_1, \theta_2, \theta_3)$ is a vector of the infinite-dimensional nuisance parameters given the non-parametric models, and the third equality follows because $\prod_{k=1}^{M} f\left(A_{v_k} \mid \overline{A}_{v_{k-1}} = \overline{0}, \overline{L}_{v_k}; \theta_3\right)$ can be equivalently expressed as the likelihood based on the data (T, Γ) given $\overline{V}_T$.

Second, we characterize Λ_k, the nuisance tangent space for θ_k, for $k = 1, 2, 3$. Assuming $f\left\{H(\psi^*); \theta_1\right\}$ and $\prod_{k=1}^{M} f\left\{L_{v_k} \mid \overline{A}_{v_{k-1}} = \overline{0}, \overline{L}_{v_{k-1}}, H(\psi^*); \theta_2\right\}$ are nonparametric, it follows from Section 4.4 of Tsiatis (2006) that the tangent space regarding θ_1 is

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

and the tangent space of θ_2 is

$$\Lambda_2 = \sum_{k=1}^{M} \left\{S\left\{\overline{V}_{v_k - 1}, L_{v_k}, H(\psi^*)\right\} \in \mathcal{R}^p :\right.$$

$$\mathbb{E}\left[S\left\{\overline{V}_{v_k-1}, L_{v_k}, H(\psi^*)\right\} \mid \overline{A}_{v_{k-1}} = \overline{0}, \overline{L}_{v_{k-1}}, H(\psi^*)\right] = 0\right\}.$$

By writing

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

it follows from Tsiatis (2006) that the tangent space of θ_3 is

$$\Lambda_3 = \left\{\int h_u(\overline{V}_u)\mathrm{d}M_T(u) : \text{for all } h_u(\overline{V}_u) \in \mathcal{R}^p\right\}.$$

Because θ_1, θ_2, and θ_3 separate out in the likelihood function, Λ_1, Λ_2 and Λ_3 are mutually orthogonal. Then, the nuisance tangent space becomes $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$, where $\oplus$ denotes a direct sum.

Third, we characterize $\Lambda^{\perp}$ using the following technical trick. Define

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

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 D, and because the tangent space for a nonparametric model is the entire Hilbert space, we obtain that $\mathcal{H} = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3^*$. Because $\Lambda^{\perp}$ must be orthogonal to $\Lambda_1 \oplus \Lambda_2$, $\Lambda^{\perp}$ consists of all elements of Λ_3^* that are orthogonal to Λ_3. It then suffices to find the projection of all elements of Λ_3^*, $\int h_u\{\overline{V}_u, H(\psi^*)\}\mathrm{d}M_T(u)$, onto $\Lambda_3^{\perp}$. To find the projection, we derive $h_u^*(\overline{V}_u)$ such that

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

Therefore, we have

$$\mathbb{E}\left(\int \left[h_u\{\overline{V}_u, H(\psi^*)\} - h_u^*(\overline{V}_u)\right]\mathrm{d}M_T(u) \times \int h_u(\overline{V}_u)\mathrm{d}M_T(u)\right) = 0, \tag{S4}$$

for any $h_u(\overline{V}_u)$. It is important to note that by Assumption 5, $M_T(t)$ is a martingale with respect to the filtration $\sigma\{\overline{V}_t, H(\psi^*)\}$. If $P_1(u)$ and $P_2(u)$ are locally bounded $\sigma\{\overline{V}_t, H(\psi^*)\}$-predictable processes, then we have the following useful result:

$$\mathbb{E}\left\{\int_0^t P_1(u)\mathrm{d}M_T(u) \int_0^t P_2(u)\mathrm{d}M_T(u)\right\} = \int_0^t P_1(u)P_2(u)\lambda_T(u \mid \overline{V}_u)R_T(u)\mathrm{d}u. \tag{S5}$$

By (S5), (S4) becomes

$$\mathbb{E}\left(\int \left[h_u\{\overline{V}_u, H(\psi^*)\} - h_u^*(\overline{V}_u)\right] h_u(\overline{V}_u)\lambda_T(u \mid \overline{V}_u)R_T(u)\mathrm{d}u\right)$$
$$= \mathbb{E}\left(\int \mathbb{E}\left(\left[h_u\{\overline{V}_u, H(\psi^*)\} - h_u^*(\overline{V}_u)\right] R_T(u) \mid \overline{V}_u\right) h_u(\overline{V}_u)\lambda_T(u \mid \overline{V}_u)\mathrm{d}u\right) = 0,$$

for any $h_u(\overline{V}_u)$. Because $h_u(\overline{V}_u)$ is arbitrary, we obtain

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

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

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

This completes the proof.

Web Appendix 1.4 *A lemma*

The following lemmas are useful for the proof of theorems.

LEMMA S1: *For any $h\{H(\psi^*), \overline{V}_u\}$, we have*

$$\mathbb{E}\int_0^\tau h\{H(\psi^*), \overline{V}_u\}\mathrm{d}M_T(u) = \mathbb{E}\int_0^\tau \big(\mathbb{E}[h\{H(\psi^*), \overline{V}_u\} \mid \overline{V}_u, T = u]$$
$$- \mathbb{E}[h\{H(\psi^*), \overline{V}_u\} \mid \overline{V}_u, T \geqslant u]\big)\lambda_T(u \mid \overline{V}_u)R_T(u)\mathrm{d}u.$$

Proof of Lemma S1.

Combining the following two results

$$\mathbb{E}\int_0^\tau h\{H(\psi^*), \overline{V}_u\}N_T(u) = \mathbb{E}\int_0^\tau \mathbb{E}[h\{H(\psi^*), \overline{V}_u\} \mid \overline{V}_u, T]\mathrm{d}N_T(u)$$
$$= \mathbb{E}\int_0^\tau \mathbb{E}[h\{H(\psi^*), \overline{V}_u\} \mid \overline{V}_u, T = u]\lambda_T(u \mid \overline{V}_u)R_T(u)\mathrm{d}u,$$

and

$$\mathbb{E}\left\{\int_0^\tau h\{H(\psi^*), \overline{V}_u\}\lambda_T(u \mid \overline{V}_u)R_T(u)\mathrm{d}u\right\}$$
$$= \mathbb{E}\left\{\int_0^\tau \mathbb{E}[h\{H(\psi^*), \overline{V}_u\} \mid \overline{V}_u, R_T(u)]\lambda_T(u \mid \overline{V}_u)R_T(u)\mathrm{d}u\right\}$$
$$= \mathbb{E}\left\{\int_0^\tau \mathbb{E}[h\{H(\psi^*), \overline{V}_u\} \mid \overline{V}_u, T \geqslant u]\lambda_T(u \mid \overline{V}_u)R_T(u)\mathrm{d}u\right\}$$

leads to Lemma S1. □

Web Appendix 1.5 *Projection onto* $\Lambda^{\perp}$

To derive $S_{\text{eff}}(\psi^*; F)$, we calculate the projection of any $B = B(F)$ onto $\Lambda^{\perp}$. Theorem S1 summarizes the result.

THEOREM S1: *For any* $B = B(F)$*, the projection of* B *onto* $\Lambda^{\perp}$ *is*

$$\prod (B \mid \Lambda^{\perp}) = \int_0^{\tau} \left[\mathbb{E}\left\{B\dot{H}_u(\psi^*) \mid \overline{V}_u, T = u\right\} - \mathbb{E}\left\{B\dot{H}_u(\psi^*) \mid \overline{V}_u, T \geqslant u\right\}\right] \times \left[\text{var}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\right\}\right]^{-1} \left[H(\psi^*) - \mathbb{E}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\right\}\right] \mathrm{d}M_T(u), \quad \text{(S7)}$$

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

Proof of Theorem S1.

For any $B = B(F)$, let

$$G = G(F) = \int_0^{\tau} \left[\mathbb{E}\left\{B\dot{H}_u(\psi^*) \mid \overline{V}_u, T = u\right\} - \mathbb{E}\left\{B\dot{H}_u(\psi^*) \mid \overline{V}_u, T \geqslant u\right\}\right] \times \left[\text{var}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\right\}\right]^{-1} \left[H(\psi^*) - \mathbb{E}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\right\}\right] \mathrm{d}M_T(u).$$

To show $\prod \left(B \mid \Lambda^{\perp}\right) = G$, it is easy to see that $G \in \Lambda^{\perp}$, so the remaining is to show that $B - G \in \Lambda$. Toward this end, we show that for any $\tilde{G} = \tilde{G}(F) = \int_0^{\tau} \tilde{c}(u, \overline{V}_u)[H(\psi^*) - \mathbb{E}\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\}]R_T(u)\mathrm{d}M_T(u) \in \Lambda^{\perp}$, $(B - G) \perp\!\!\!\perp \tilde{G}$ or $\mathbb{E}\{(B - G)\tilde{G}\} = 0$. We now verify $\mathbb{E}(B\tilde{G}) = \mathbb{E}(G\tilde{G})$ by the following calculation.

First, by (S5), we calculate

$$\begin{aligned}
\mathbb{E}(G\tilde{G}) &= \mathbb{E}\int_0^{\tau} \tilde{c}(u, \overline{V}_u)[\mathbb{E}\{B\dot{H}_u(\psi^*) \mid \overline{V}_u, T = u\} - \mathbb{E}\{B\dot{H}_u(\psi^*) \mid \overline{V}_u, T \geqslant u\}] \\
&\quad \times [\text{var}\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\}]^{-1}[H(\psi^*) - \mathbb{E}\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\}]^2 \\
&\quad \times \lambda_T(u \mid \overline{V}_u)R_T(u)\mathrm{d}u \\
&= \mathbb{E}\int_0^{\tau} \tilde{c}(u, \overline{V}_u)[\mathbb{E}\{B\dot{H}_u(\psi^*) \mid \overline{V}_u, T = u\} - \mathbb{E}\{B\dot{H}_u(\psi^*) \mid \overline{V}_u, T \geqslant u\}] \\
&\quad \times \lambda_T(u \mid \overline{V}_u)R_T(u)\mathrm{d}u. \qquad \text{(S8)}
\end{aligned}$$

Second, we calculate

$$\begin{aligned}
\mathbb{E}(B\tilde{G}) &= \mathbb{E}\int_0^\tau \tilde{c}(u,\overline{V}_u)B[H(\psi^*)-\mathbb{E}\{H(\psi^*)\mid \overline{L}_u, T\geqslant u\}]\mathrm{d}M_T(u)\\
&= \mathbb{E}\int_0^\tau \tilde{c}(u,\overline{V}_u)B\dot{H}_u(\psi^*)\mathrm{d}N_T(u)\\
&\quad -\mathbb{E}\int_0^\tau \tilde{c}(u,\overline{V}_u)B\dot{H}_u(\psi^*)\lambda_T(u\mid \overline{V}_u)R_T(u)\mathrm{d}u\\
&= \mathbb{E}\int_0^\tau \tilde{c}(u,\overline{V}_u)[\mathbb{E}\{B\dot{H}_u(\psi^*)\mid \overline{V}_u, T=u\}-\mathbb{E}\{B\dot{H}_u(\psi^*)\mid \overline{V}_u, T\geqslant u\}]\\
&\quad \times\lambda_T(u\mid \overline{V}_u)R_T(u)\mathrm{d}u, \qquad \text{(S9)}
\end{aligned}$$

where the last equality follows by Lemma S1.

Therefore, by (S8) and (S9), $\mathbb{E}(B\tilde{G})=\mathbb{E}(G\tilde{G})$ for any $\tilde{G}\in\Lambda^\perp$, proving (S7). □

Web Appendix 1.6 *Proof of Theorem 2*

The SES is $S^*_{\text{eff}}(\psi^*)=\prod\left(S_\psi\mid\Lambda^\perp\right)$. By Theorem S1, we have

$$\begin{aligned}
S^*_{\text{eff}}(\psi^*) &= \int_0^\tau [\mathbb{E}\{S_\psi\dot{H}_u(\psi^*)\mid \overline{V}_u, T=u\}-\mathbb{E}\{S_\psi\dot{H}_u(\psi^*)\mid \overline{V}_u, T\geqslant u\}]\\
&\quad \times[\mathrm{var}\{H(\psi^*)\mid \overline{V}_u, T\geqslant u\}]^{-1}[H(\psi^*)-\mathbb{E}\{H(\psi^*)\mid \overline{V}_u, T\geqslant u\}]\mathrm{d}M_T(u)\\
&= -\int_0^\tau [\mathbb{E}\{\partial\dot{H}_u(\psi^*)/\partial\psi\mid \overline{V}_u, T=u\}-\mathbb{E}\{\partial\dot{H}_u(\psi^*)/\partial\psi\mid \overline{V}_u, T\geqslant u\}]\\
&\quad \times[\mathrm{var}\{H(\psi^*)\mid \overline{V}_u, T\geqslant u\}]^{-1}[H(\psi^*)-\mathbb{E}\{H(\psi^*)\mid \overline{V}_u, T\geqslant u\}]\mathrm{d}M_T(u)\\
&= -\int_0^\tau \mathbb{E}\{\partial\dot{H}_u(\psi^*)/\partial\psi\mid \overline{V}_u, T=u\}[\mathrm{var}\{H(\psi^*)\mid \overline{V}_u, T\geqslant u\}]^{-1}\\
&\quad \times H(\psi^*)-\mathbb{E}\{H(\psi^*)\mid \overline{V}_u, T\geqslant u\}]\mathrm{d}M_T(u),
\end{aligned}$$

where the last equality follows by using the generalized information equality: because $\dot{H}_u(\psi^*)=H(\psi^*)-\mathbb{E}\{H(\psi^*)\mid \overline{V}_u, T\geqslant u\}$, we have $\mathbb{E}\{\dot{H}_u(\psi^*)\mid \overline{V}_u, T\geqslant u\}=0$. Take the derivative of ψ at both sides, we have $\mathbb{E}\{S_\psi\dot{H}_u(\psi^*)\mid \overline{V}_u, T\geqslant u\}+\mathbb{E}\{\partial\dot{H}_u(\psi^*)/\partial\psi\mid \overline{V}_u, T\geqslant u\}=0$, or equivalently $\mathbb{E}\{S_\psi\dot{H}_u(\psi^*)\mid \overline{V}_u, T\geqslant u\}=-\mathbb{E}\{\partial\dot{H}_u(\psi^*)/\partial\psi\mid \overline{V}_u, T\geqslant u\}$. Similarly, noticing $\mathbb{E}\{H(\psi^*)\mid \overline{V}_u, T\geqslant u\}=\mathbb{E}\{H(\psi^*)\mid \overline{V}_u, T=u\}$, we have $\mathbb{E}\{\dot{H}_u(\psi^*)\mid \overline{V}_u, T=u\}=0$. Take the derivative of ψ at both sides, we have $\mathbb{E}\{S_\psi\dot{H}_u(\psi^*)\mid \overline{V}_u, T=u\}+\mathbb{E}\{\partial\dot{H}_u(\psi^*)/\partial\psi\mid \overline{V}_u, T=$

$u\} = 0$, or equivalently $\mathbb{E}\{S_\psi \dot{H}_u(\psi^*) \mid \overline{V}_u, T = u\} = -\mathbb{E}\{\partial \dot{H}_u(\psi^*)/\partial\psi \mid \overline{V}_u, T = u\}$. Ignoring the negative sign, the result in Theorem 2 follows.

Web Appendix 1.7 *Proof of Theorem 3*

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

First, if $\lambda_T(t \mid \overline{V}_t)$ is correctly specified, under Assumption 5, $M_T(t)$ is a martingale with respect to the filtration $\sigma\{\overline{V}_t, H(\psi^*)\}$. Because $c(u, \overline{V}_u)[H(\psi^*) - \mathbb{E}\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\}]$ is a $\sigma\{\overline{V}_t, H(\psi^*)\}$-predictable process, $\int_0^t c(u, \overline{V}_u)[H(\psi^*) - \mathbb{E}\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\}]\mathrm{d}M_T(u)$ is a martingale for $t \geqslant 0$. Therefore, $\mathbb{E}\{G(\psi^*; F, c)\} = 0$.

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

$$
\begin{aligned}
& \mathbb{E}\int c(u, \overline{V}_u)\left[H(\psi^*) - \mathbb{E}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u; \beta^*\right\}\right]\left\{\mathrm{d}N_T(u) - \lambda_T^*(u \mid \overline{V}_u)R_T(u)\mathrm{d}u\right\} \\
= \; & \mathbb{E}\int c(u, \overline{V}_u)\left[H(\psi^*) - \mathbb{E}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u; \beta^*\right\}\right]\left\{\mathrm{d}N_T(u) - \lambda_T(u \mid \overline{V}_u)R_T(u)\mathrm{d}u\right\} \\
& + \mathbb{E}\int c(u, \overline{V}_u)\left[H(\psi^*) - \mathbb{E}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u; \beta^*\right\}\right]\left\{\lambda_T(u \mid \overline{V}_u) - \lambda_T^*(u \mid \overline{V}_u)\right\}R_T(u)\mathrm{d}u \\
= \; & 0 + \mathbb{E}\int c(u, \overline{V}_u)E\left(\left[H(\psi^*) - \mathbb{E}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u; \beta^*\right\}\right] \mid \overline{V}_u, T \geqslant u\right) \qquad \text{(S10)} \\
& \times \left\{\lambda_T(u \mid \overline{V}_u) - \lambda_T^*(u \mid \overline{V}_u)\right\}R_T(u)\mathrm{d}u \\
= \; & 0 + \mathbb{E}\int c(u, \overline{V}_u) \times 0 \times \left\{\lambda_T(u \mid \overline{V}_u) - \lambda_T^*(u \mid \overline{V}_u)\right\}R_T(u)\mathrm{d}u = 0, \qquad \text{(S11)}
\end{aligned}
$$

where zero in (S10) follows because $\mathrm{d}M_T(u) = \mathrm{d}N_T(u) - \lambda_T(u \mid \overline{V}_u)\mathrm{d}u$ is a martingale with respect to the filtration $\sigma\{\overline{V}_t, H(\psi^*)\}$, and zero in (S11) follows because $\mathbb{E}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\right\}$ is correctly specified and therefore, $\mathbb{E}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u; \beta^*\right\} = \mathbb{E}\left\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\right\}$.

Web Appendix 1.8 *Proof of Theorem 4*

The proposed estimator $\widehat{\psi}$ depends on two nuisance models: (i) $\mathbb{E}\{H(\psi^*) \mid \overline{V}_u, T \geqslant u; \beta\}$ indexed by β and (ii) the proportional hazards model for the treatment process, denoted by M_T. Let $\widehat{\beta}$

and $\widehat{M}_T$ be the estimates of β and M_T under the specified parametric and semiparametric models, respectively. To reflect that the estimating function depends on the nuisance parameters, we denote $G(\psi, \beta, M_T; F) = \int c(u, \overline{V}_u)[H(\psi) - \mathbb{E}\{H(\psi) \mid \overline{V}_u, T \geqslant u; \beta\}]\mathrm{d}M_T(u)$. Denote the probability limits of $\widehat{\beta}$ and $\widehat{M}_T$ as β^* and M_T^*, respectively.

Proof of the double robustness.

The key implications of Assumption 6 are the following. First, By Assumptions 4 and 6, $H(\psi^*) \sim Y^{(\infty)}$, given $(\overline{L}_t, T \geqslant t, C \geqslant t)$. Second, under Assumption 6, the hazard of treatment initiation in (7) is equal to $\lim_{h\to 0} h^{-1}P(t \leqslant T < t + h, \Gamma = 1 \mid \overline{V}_t, T \geqslant t, C \geqslant t)$. Thus, we can follow the same steps in Section Web Appendix 1.7 to show the double robustness by changing the conditioning set $\{\overline{V}_t, T \geqslant t\}$ to $\{\overline{V}_t, T \geqslant t, C \geqslant t\}$. □

Proof of the optimality.

Denote the class of estimating functions for ψ as $\mathcal{G} = \{G_C(\psi, g; F) = \delta_C g(\overline{V}_\tau)G(\psi; F) : \text{for any } g\}$, where $G(\psi; F)$ is the SES of ψ in the absence of censoring. For simplicity, let $X^{\otimes 2}$ denote XX^{T}. The following lemma gives a condition under which g^* is optimal among all choices of g in $\mathcal{G}$.

LEMMA S2: *If g^* satisfies*

$$\mathbb{E}\left\{\partial G_C(\psi, g; F)\partial\psi|_{\psi=\psi^*}\right\} = \mathbb{E}\left\{G_C(\psi^*, g; F)G_C(\psi^*, g^*; F)^{\mathrm{T}}\right\}, \tag{S12}$$

for any g, then the solution to the estimating equation constructed based on $G_C(\psi, g^*; F)$ for ψ^* achieves the smallest asymptotic variance.

We show that $g^*(\overline{V}_\tau) = 1$ satisfies the optimality criterion (S12). Recall that $G(\psi; F) = \int_0^\tau c_{\text{eff}}(u, \overline{V}_u)\dot{H}_u(\psi^*)R_T(u)\mathrm{d}M_T(u)$ with

$$c_{\text{eff}}(u, \overline{V}_u) = [\mathbb{E}\{\partial\dot{H}_u(\psi^*)/\partial\psi \mid \overline{V}_u, T = u\} - \mathbb{E}\{\partial\dot{H}_u(\psi^*)/\partial\psi \mid \overline{V}_u, T \geqslant u\}] \times [\mathrm{var}\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\}]^{-1}. \tag{S13}$$

For any g, define $\kappa(\overline{V}_u; g) = \mathbb{E}\{K_C(\overline{V}_\tau)g(\overline{V}_\tau) \mid \overline{V}_u\}$. The left hand side of (S12) is

$$\begin{aligned}
& \mathbb{E}\left\{ \partial G_C(\psi, g; F)\partial\psi|_{\psi=\psi^*} \right\} \\
= \; & \mathbb{E}\left[\delta_C g(\overline{V}_\tau) \left\{\partial G(\psi; F)/\partial\psi\right\}\big|_{\psi=\psi^*} \right] \\
= \; & \mathbb{E}\left[K_C(\overline{V}_\tau) g(\overline{V}_\tau) \left\{\partial G(\psi; F)/\partial\psi\right\}\big|_{\psi=\psi^*} \right] \\
= \; & \mathbb{E}\left[\kappa(\overline{V}_\tau; g) \int_0^\tau c_{\text{eff}}(\overline{V}_u) \left\{\partial \dot{H}_u(\psi^*)/\partial\psi\right\} R_T(u)\mathrm{d}M_T(u) \right]. \qquad \text{(S14)}
\end{aligned}$$

Similarly, the right hand side of (S12) with $g^*(\overline{V}_\tau) = 1$ is

$$\begin{aligned}
\mathbb{E}\left\{G_C(\psi^*, g; F)^{\otimes 2}\right\} &= \mathbb{E}\left\{K_C(\overline{V}_\tau)g(\overline{V}_\tau)G(\psi^*; F)^{\otimes 2}\right\} \\
&= \mathbb{E}\left[\kappa(\overline{V}_\tau; g) \int_0^\tau c_{\text{eff}}(\overline{V}_u)^{\otimes 2} \{\dot{H}_u(\psi^*)\}^2 \lambda_T(u \mid \overline{V}_u) R_T(u)\mathrm{d}u \right] \qquad \text{(S15)}
\end{aligned}$$

We now show (S12) by induction. Let $\Delta > 0$ be a small increment. We start with (S14), which becomes

$$\begin{aligned}
& \mathbb{E}\left[\kappa(\overline{V}_\tau; g) \left(\int_0^{\tau-\Delta} + \int_{\tau-\Delta}^\tau \right) c_{\text{eff}}(u, \overline{V}_u) \left\{\partial \dot{H}_u(\psi^*)/\partial\psi\right\} R_T(u)\mathrm{d}M_T(u) \right] \\
= \; & \mathbb{E}\left[\kappa(\overline{V}_{\tau-\Delta}; g) \int_0^{\tau-\Delta} c_{\text{eff}}(u, \overline{V}_u) \left\{\partial \dot{H}_u(\psi^*)/\partial\psi\right\} R_T(u)\mathrm{d}M_T(u) \right] \\
& + \mathbb{E}\left[\kappa(\overline{V}_\tau; g) \int_{\tau-\Delta}^\tau c_{\text{eff}}(u, \overline{V}_u) \left\{\partial \dot{H}_u(\psi^*)/\partial\psi\right\} R_T(u)\mathrm{d}M_T(u) \right]. \qquad \text{(S16)}
\end{aligned}$$

Also, (S15) becomes

$$\begin{aligned}
& \mathbb{E}\left[\kappa(\overline{V}_\tau; g) \left(\int_0^{\tau-\Delta} + \int_{\tau-\Delta}^\tau \right) c_{\text{eff}}(u, \overline{V}_u)^{\otimes 2} \{\dot{H}_u(\psi^*)\}^2 \lambda_T(u \mid \overline{V}_u) R_T(u)\mathrm{d}u \right] \\
= \; & \mathbb{E}\left[\kappa(\overline{V}_{\tau-\Delta}; g) \int_0^{\tau-\Delta} c_{\text{eff}}(u, \overline{V}_u)^{\otimes 2} \{\dot{H}_u(\psi^*)\}^2 \lambda_T(u \mid \overline{V}_u) R_T(u)\mathrm{d}u \right] \\
& + \mathbb{E}\left[\kappa(\overline{V}_\tau; g) \int_{\tau-\Delta}^\tau c_{\text{eff}}(u, \overline{V}_u)^{\otimes 2} \{\dot{H}_u(\psi^*)\}^2 \lambda_T(u \mid \overline{V}_u) R_T(u)\mathrm{d}u \right]. \qquad \text{(S17)}
\end{aligned}$$

It suffices to show that the term in (S16) equals the term in (S17). By Lemma S1, the term in (S16) is

$$\begin{aligned}
& \mathbb{E}\left[\kappa(\overline{V}_\tau; g) \int_{\tau-\Delta}^\tau c_{\text{eff}}(u, \overline{V}_u) \left\{\partial \dot{H}_u(\psi^*)/\partial\psi\right\} R_T(u)\mathrm{d}M_T(u) \right] \\
= \; & \mathbb{E}\kappa(\overline{V}_\tau; g) \int_{\tau-\Delta}^\tau c_{\text{eff}}(u, \overline{V}_u) \left[\mathbb{E}\left\{\partial \dot{H}_u(\psi^*)/\partial\psi \mid \overline{V}_u, T = u\right\} \right. \\
& \left. - \mathbb{E}\left\{\partial \dot{H}_u(\psi^*)/\partial\psi \mid \overline{V}_u, T = u\right\} \right] \lambda_T(u \mid \overline{V}_u) R_T(u)\mathrm{d}u,
\end{aligned}$$

and the term in (S17) is

$$\begin{aligned}
& \mathbb{E}\left[\kappa(\overline{V}_\tau; g)\int_{\tau-\Delta}^{\tau} c_{\text{eff}}(u, \overline{V}_u)^{\otimes 2}\{\dot{H}_u(\psi^*)\}^2 \lambda_T(u \mid \overline{V}_u) R_T(u)\mathrm{d}u\right] \\
= \; & \mathbb{E}\left\{\kappa(\overline{V}_\tau; g)\int_{\tau-\Delta}^{\tau} c_{\text{eff}}(u, \overline{V}_u)^{\otimes 2}\mathbb{E}[\{\dot{H}_u(\psi^*)\}^2 \mid \overline{V}_u, R_T(u)]\lambda_T(u \mid \overline{V}_u) R_T(u)\mathrm{d}u\right\} \\
= \; & \mathbb{E}\kappa(\overline{V}_\tau; g)\int_{\tau-\Delta}^{\tau} c_{\text{eff}}(u, \overline{V}_u)\left[\mathbb{E}\left\{\partial\dot{H}_u(\psi^*)/\partial\psi \mid \overline{V}_u, T=u\right\}\right. \\
& \left. - \mathbb{E}\left\{\partial\dot{H}_u(\psi^*)/\partial\psi \mid \overline{V}_u, T=u\right\}\right]\lambda_T(u \mid \overline{V}_u) R_T(u)\mathrm{d}u,
\end{aligned}$$

and therefore the two terms are equal. Then, (S12) follows by induction. This completes the proof. □

Web Appendix 2. Asymptotic distribution of $\widehat{\psi}$

Let $\Phi(\psi, \beta, M_T; F) = \delta_C G(\psi, \beta, M_T; F)$, for $\psi \in \Theta$, $\beta \in \mathcal{B}$, and $M_T \in \mathcal{F}_T$, where Θ and $\mathcal{B}$ are compact sets in the Euclidean space and $\mathcal{F}_T$ contains M_T with a bounded L_2 norm. Then, the proposed estimator $\widehat{\psi}$ solves

$$\mathbb{P}_n\left\{\Phi(\psi, \widehat{\beta}, \widehat{M_T}; F)\right\} = 0, \tag{S18}$$

for ψ. We present the asymptotic distribution of the proposed estimator $\widehat{\psi}$ solving equation (S18). Let $\mathbb{P}$ denote the true data generating distribution of F, and for any $g(F)$, let $\mathbb{P}\{g(F)\} = \int g(f)\mathrm{d}\mathbb{P}(f)$ and let $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - \mathbb{P})$. We define

$$\begin{aligned}
J_1(\beta) &= \mathbb{P}\left\{\Phi(\psi^*, \beta, M_T^*; F)\right\}, \\
J_2(M_T) &= \mathbb{P}\left\{\Phi(\psi^*, \beta^*, M_T; F)\right\}, \\
J(\beta, M_T) &= \mathbb{P}\left\{\Phi(\psi^*, \beta, M_T; F)\right\}.
\end{aligned}$$

Similar to Yang and Lok (2016), we impose the regularity conditions from the empirical process literature (van der Vaart and Wellner, 1996).

ASSUMPTION S1: (i) For any sequence of $\{\psi_n\} \in \Theta$, if $||\mathbb{P}\Phi(\psi_n, \beta^*, M_T^*; F)|| \to 0$, then $||\psi_n - \psi^*|| \to 0$.

(ii) $\Phi(\psi, \beta, M_T; F)$ and $\partial\Phi(\psi, \beta, M_T; F)/\partial\psi$ are weak Glivenko-Cantelli classes; i.e.,

$$\begin{aligned} \sup_{\psi\in\Theta,\beta\in\mathcal{B},M_T\in\mathcal{F}_T} |\mathbb{P}_n - \mathbb{P}|\{\Phi(\psi, \beta, M_T; F)\} &= o_{\mathbb{P}}(1), \\ \sup_{\psi\in\Theta,\beta\in\mathcal{B},M_T\in\mathcal{F}_T} |\mathbb{P}_n - \mathbb{P}| \left\{ \frac{\partial\Phi(\psi, \beta, M_T; F)}{\partial\psi} \right\} &= o_{\mathbb{P}}(1). \end{aligned}$$

(iii) $\Phi(\psi, \beta, M_T; F)$ and $\partial\Phi(\psi, \beta, M_T; F)/\partial\psi$ are Donsker classes; i.e.,

$$\begin{aligned} \mathbb{G}_n\{\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F)\} &= \mathbb{G}_n\{\Phi(\psi^*, \beta^*, M_T^*; F)\} + o_{\mathbb{P}}(1), \\ \mathbb{G}_n \left\{ \frac{\partial\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F)}{\partial\psi} \right\} &= \mathbb{G}_n \left\{ \frac{\partial\Phi(\psi^*, \beta^*, M_T^*; F)}{\partial\psi} \right\} + o_{\mathbb{P}}(1). \end{aligned}$$

(iv) Assume that

$$\begin{aligned} \mathbb{P}\left\{ ||\Phi(\psi^*, \widehat{\beta}, \widehat{M}_T; F) - \Phi(\psi^*, \beta^*, M_T^*; F)|| \right\} &= o_{\mathbb{P}}(1), \\ \mathbb{P}\left\{ ||\frac{\partial}{\partial\psi}\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F) - \frac{\partial}{\partial\psi}\Phi(\psi^*, \beta^*, M_T^*; F)|| \right\} &= o_{\mathbb{P}}(1). \end{aligned}$$

(v) $A(\psi^*, \beta^*, M_T^*) = \mathbb{P}\{\partial\Phi(\psi^*, \beta^*, M_T^*; F)/\partial\psi\}$ is invertible.

(vi) Assume that

$$J(\widehat{\beta}, \widehat{M}_T) - J(\beta^*, M_T^*) = J_1(\widehat{\beta}) - J_1(\beta^*) + J_2(\widehat{M}_T) - J_2(M_T^*) + o_{\mathbb{P}}(n^{-1/2}),$$

and that $J_1(\widehat{\beta})$ and $J_2(\widehat{M}_T)$ are regular asymptotically linear with influence functions $\Phi_1(\psi^*, \beta^*, M_T^*; F)$ and $\Phi_2(\psi^*, \beta^*, M_T^*; F)$, respectively.

We discuss the implications of these conditions. First, Assumption S1 (i) is an identification condition that ensures the uniqueness of the solution ψ^* to the estimating equation $\mathbb{P}\Phi(\psi^*, \beta^*, M_T^*; F) = 0$. Assumption S1 (i) and (ii) are standard for the consistency of $\widehat{\psi}$. The Donsker class condition requires that the nuisance models should not be too complex. Assumption S1 (iii) is a standard condition for the empirical processes. We refer the interested readers to Section 4.2 of Kennedy (2016) for a thorough discussion of Donsker classes of functions. Second, Assumption S1 (iv) states that $\widehat{\beta}$ and $\widehat{M}_T$ have probability limits β^* and M_T^*, and that the double robustness condition in Theorem 4 holds. Third, Assumption S1 (vi) holds for smooth functionals of parametric or semiparametric efficient estimators under specified models. Therefore, this assumption would hold

under mild regularity conditions if $\widehat{\beta}$ and $\widehat{M}_T$ are the parametric and semiparametric maximum likelihood estimators under specified models.

THEOREM S2: *Under the continuous-time SNMM (6) and Assumptions 5 and S1,* $\widehat{\psi}$ *is consistent for* ψ^* *and satisfies*

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

where $\Sigma_{\psi\psi} = \text{var}\{\widetilde{\Phi}(\psi^*, \beta^*, M_T^*; F)\}$, $\widetilde{\Phi}(\psi^*, \beta^*, M_T^*; F) = \{A(\psi^*, \beta^*, M_T^*)\}^{-1}\widetilde{B}(\psi^*, \beta^*, M_T^*; F)$, $A(\psi^*, \beta^*, M_T^*)$ *is defined in Assumption S1 (iii),*

$$\widetilde{B}(\psi^*, \beta^*, M_T^*; F) = \Phi(\psi^*, \beta^*, M_T^*; F) + \Phi_1(\psi^*, \beta^*, M_T^*; F) + \Phi_2(\psi^*, \beta^*, M_T^*; F), \qquad \text{(S19)}$$

and $\Phi_1(\psi^*, \beta^*, M_T^*; F)$ *and* $\Phi_2(\psi^*, \beta^*, M_T^*; F)$ *are defined in Assumption S1 (iv).*

Proof of Theorem S2.

We assume the double robustness condition holds; i.e., either the potential outcome mean model or the model for the treatment process is correctly specified. By Assumption S1 (i), we have

$$\begin{aligned}
||\mathbb{P}\{\Phi(\widehat{\psi}, \beta^*, M_T^*; F)\}|| &\leqslant ||\mathbb{P}\{\Phi(\widehat{\psi}, \beta^*, M_T^*; F)\} - \mathbb{P}\{\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F)\}|| + ||\mathbb{P}\{\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F)\}|| \\
&= ||\mathbb{P}\{\Phi(\widehat{\psi}, \beta^*, M_T^*; F)\} - \mathbb{P}\{\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F)\}|| \\
&\quad + ||\mathbb{P}\{\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F)\} - \mathbb{P}_n\{\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F)\}|| \\
&= ||\mathbb{P}\{\Phi(\widehat{\psi}, \beta^*, M_T^*; F)\} - \mathbb{P}\{\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F)\}|| \\
&\quad + \sup_{\psi\in\Theta, \beta\in\mathcal{B}, M_T\in\mathcal{F}_T} |\mathbb{P}_n - \mathbb{P}|\{\Phi(\psi, \beta, M_T; F)\} \\
&= o_{\mathbb{P}}(1).
\end{aligned}$$

Together with Assumption S1 (ii) leads to the consistency of $\widehat{\psi}$.

Applying the Taylor expansion of $\mathbb{P}_n\left\{\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F)\right\} = 0$ around ψ^* leads to

$$0 = \mathbb{P}_n\left\{\Phi(\widehat{\psi}, \widehat{\beta}, \widehat{M}_T; F)\right\}$$

$$= \mathbb{P}_n\left\{\Phi(\psi^*, \widehat{\beta}, \widehat{M}_T; F)\right\} + \mathbb{P}_n\left\{\frac{\partial\Phi(\widetilde{\psi}, \widehat{\beta}, \widehat{M}_T; F)}{\partial\psi^{\mathrm{T}}}\right\}(\widehat{\psi} - \psi^*),$$

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

Under Assumption S1 (iii) and (iv),

$$(\mathbb{P}_n - \mathbb{P})\left\{\frac{\partial\Phi(\widetilde{\psi}, \widehat{\beta}, \widehat{M}_T; F)}{\partial\psi^{\mathrm{T}}}\right\} = (\mathbb{P}_n - \mathbb{P})\left\{\frac{\partial\Phi(\psi^*, \beta^*, M_T^*; F)}{\partial\psi^{\mathrm{T}}}\right\} = o_{\mathbb{P}}(n^{-1/2}),$$

and therefore,

$$\begin{aligned}\mathbb{P}_n\left\{\frac{\partial\Phi(\widetilde{\psi}, \widehat{\beta}, \widehat{M}_T; F)}{\partial\psi^{\mathrm{T}}}\right\} &= \mathbb{P}\left\{\frac{\partial\Phi(\widetilde{\psi}, \widehat{\beta}, \widehat{M}_T; F)}{\partial\psi^{\mathrm{T}}}\right\} + o_{\mathbb{P}}(n^{-1/2}) \\ &= A(\psi^*, \beta^*, M_T^*) + o_{\mathbb{P}}(n^{-1/2}).\end{aligned}$$

We then have

$$n^{1/2}(\widehat{\psi} - \psi^*) = \{A(\psi^*, \beta^*, M_T^*)\}^{-1} n^{1/2}\mathbb{P}_n\left\{\Phi(\psi^*, \widehat{\beta}, \widehat{M}_T; F)\right\} + o_{\mathbb{P}}(1). \tag{S20}$$

Based on the double robustness, we have

$$\mathbb{P}\{\Phi(\psi^*, \beta^*, M_T^*; F)\} = 0. \tag{S21}$$

To express (S20) further, based on (S21), we have

$$\begin{aligned}\mathbb{P}_n\Phi(\psi^*, \widehat{\beta}, \widehat{M}_T; F) &= (\mathbb{P}_n - \mathbb{P})\Phi(\psi^*, \widehat{\beta}, \widehat{M}_T; F) \\ &\quad + \mathbb{P}\left\{\Phi(\psi^*, \widehat{\beta}, \widehat{M}_T; F) - \Phi(\psi^*, \beta^*, M_T^*; F)\right\} + \mathbb{P}\Phi(\psi^*, \beta^*, M_T^*; F).\end{aligned} \tag{S22}$$

By Assumption S1 (iii) and (iv), the first term in (S6) becomes

$$\begin{aligned}(\mathbb{P}_n - \mathbb{P})\Phi(\psi^*, \widehat{\beta}, \widehat{M}_T; F) &= (\mathbb{P}_n - \mathbb{P})\Phi(\psi^*, \beta^*, M_T^*; F) + o_{\mathbb{P}}(n^{-1/2}) \\ &= \mathbb{P}_n\Phi(\psi^*, \beta^*, M_T^*; F) + o_{\mathbb{P}}(n^{-1/2}).\end{aligned} \tag{S23}$$

By Assumption S1 (vi), the second term in (S6) becomes

$$\begin{aligned}&\mathbb{P}\left\{\Phi(\psi^*, \widehat{\beta}, \widehat{M}_T; F) - \Phi(\psi^*, \beta^*, M_T^*; F)\right\} \\ =\ & J(\widehat{\beta}, \widehat{M}_T) - J(\beta^*, M_T^*) + o_{\mathbb{P}}(n^{-1/2}) \\ =\ & J_1(\widehat{\beta}) - J_1(\beta^*) + J_2(\widehat{M}_T) - J_2(M_T^*) + o_{\mathbb{P}}(n^{-1/2}) \\ =\ & \mathbb{P}_n\Phi_1(\psi^*, \beta^*, M_T^*; F) + \mathbb{P}_n\Phi_2(\psi^*, \beta^*, M_T^*; F).\end{aligned} \tag{S24}$$

Algorithm 1 Algorithm 1 for generating T according to a time-dependent proportional hazards model

Step 1. Set $k = 1$.

Step 2. Generate a temporary time to treatment initiation, $T_{\text{temp},k}$, compatible with the hazard function for the time interval $[t_k, t_{k+1})$, using the method of Bender et al. (2005); i.e., generate $u \sim$Uniform$[0, 1]$ and let $T_{\text{temp},k} = -\log(1-u)/\{\lambda_{T,0}\exp(\alpha_1^* L_{TI} + \alpha_2^* L_{TD,t_k})\}$.

If $T_{\text{temp},k}$ is contained within the first time interval $[0, t_{k+1} - t_k)$, then set $T = T_{\text{temp},k} + t_k$;

else if $T_{\text{temp},k}$ is not contained within the interval $[0, t_{k+1} - t_k)$, increase k by 1 and move to the beginning of Step 2.

Combining (S21)–(S24) leads to

$$\mathbb{P}_n \Phi(\psi^*, \widehat{\beta}, \widehat{M}_T; F) = \mathbb{P}_n\{\widetilde{B}(\psi^*, \beta^*, M_T^*; F)\},$$

where

$$\begin{aligned}\widetilde{B}(\psi^*, \beta^*, M_T^*; F) &= \Phi(\psi^*, \beta^*, M_T^*; F) + \Phi_1(\psi^*, \beta^*, M_T^*; F) \\ &\quad + \Phi_2(\psi^*, \beta^*, M_T^*; F) + \Phi_3(\psi^*, \beta^*, M_T^*; F).\end{aligned}$$

As a result,

$$n^{1/2}(\widehat{\psi} - \psi^*) = n^{1/2}\mathbb{P}_n \widetilde{\Phi}(\psi^*, \beta^*, M_T^*; F) + o_{\mathbb{P}}(1), \tag{S25}$$

where

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

This completes the proof. □

Web Appendix 3. Details for the simulation study

First, Algorithm 1 specifies the steps for generating T according to a time-dependent proportional hazards model.

Second, we describe the nuisance models and their estimation. For $c(u, \overline{V}_u)$, we approximate

$\mathbb{E}\{(1,T)^{\mathrm{T}}(\tau-T)^{+} \mid \overline{V}_u, T \geqslant u\}$ by $\widehat{\mathrm{pr}}(T \leqslant \tau \mid \overline{V}_u, T \geqslant u) \times \widehat{\mathbb{E}}\{(1,T)^{\mathrm{T}}(\tau-T) \mid \overline{V}_u, u \leqslant T \leqslant \tau\}$. We describe the details for fitting below:

(a) $\widehat{\mathrm{pr}}(T \leqslant \tau \mid \overline{V}_u, T \geqslant u)$ is the predicted value from a logistic regression model of $I(T \leqslant \tau)$ against u, L_{TI}, $L_{TD,u}$, and all interactions of these terms, restricted to subjects with $T \geqslant u$.

(b) $\widehat{\mathbb{E}}(\tau - T \mid \overline{V}_u, u \leqslant T \leqslant \tau)$ is the predicted value from a linear regression model of $\tau - T$ against u, L_{TI}, $L_{TD,u}$, and all interactions of these terms, restricted to subjects with $u \leqslant T \leqslant \tau$.

(c) $\widehat{\mathbb{E}}\{T(\tau-T) \mid \overline{V}_u, u \leqslant T \leqslant \tau\}$ is the predicted value from a linear regression model of $T(\tau-T)$ against u, L_{TI}, $L_{TD,u}$, and all interactions of these terms, restricted to subjects with $u \leqslant T \leqslant \tau$.

(d) $\mathbb{E}\{H(\widehat{\psi}_p) \mid \overline{V}_u, T \geqslant u; \widehat{\beta}\}$ by a linear regression model of $H(\widehat{\psi}_p)$ against u, L_{TI}, $L_{TD,u}$, and all interactions of these terms, restricted to subjects with $T \geqslant u$.

Tables S1 and S2 show the simulation results with $n = 2000$ in Setting I and Setting II, respectively. The same discussion applies to the estimators considered in the simulation exercise. Moreover, for the proposed estimators, the bias remains small and the standard errors decreases as n increases.

[Table 1 about here.]

[Table 2 about here.]

Web Appendix 4. Sensitivity analysis in the real data application

In the sensitivity analysis, we assume an elaborated SNMM with possible treatment effect modifiers:

$$\gamma(\overline{V}_u; \psi^*) = (\psi_1^* + \psi_2^* t + \psi_3^* \mathrm{male} + \psi_4^* \mathrm{age} + \psi_5^* \mathrm{white} + \psi_6^* \mathrm{injdrug} + \psi_7^* \mathrm{CD4}_t + \psi_8^* \mathrm{lvl}_t)(\tau - t)^{+}, \quad \text{(S26)}$$

where male is the indicator of being male, age is age at infection, white is the indicator of being white and non-hispanic, injdrug is the indicator of being an injection drug user, $\mathrm{CD4}_t$ and lvl_t are time-varying CD4 and log viral load at t. Table S3 shows the results for the sensitivity analysis

with SNMM (S26). Consistent with the primary result in the main text, the estimate of ψ_2^* is -0.36 that shows that earlier HAART initiation is better in increasing CD4 counts. However, none of the estimates of the SNMM parameters is significant, probably because of the increased number of parameters.

[Table 3 about here.]

References

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Stat Med* **24,** 1713–1723.

Bickel, P. J., Klaassen, C., Ritov, Y., and Wellner, J. (1993). *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research*, pages 141–167. Springer.

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

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Emprical Processes: With Applications to Statistics*. New York: Springer.

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

Received . Revised . Accepted .

Table S1: Simulation results in Setting I without censoring based on $1,000$ simulated datasets: the Monte Carlo bias, standard error, root mean square error of the estimators, and coverage rate of 95% confidence intervals.

		Bias ($\times 10^2$)		SE ($\times 10^2$)		rMSE ($\times 10^2$)		CR ($\times 10^2$)	
($n = 2000$)	Method	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*
Scenario (i): model for M_T ($\checkmark$)									
Model for POM ($\times$)	$\widehat{\psi}_p$	0.2	-0.1	3.4	6.2	3.4	6.2	95.9	96.0
Model for POM ($\checkmark$)	$\widehat{\psi}_{\text{cont},1}$	0.1	0.1	3.3	5.8	3.3	5.8	95.2	95.4
	$\widehat{\psi}_{\text{cont},2}$	0.1	0.1	3.2	5.6	3.2	5.6	95.1	95.6
–	$\widehat{\psi}_{\text{disc},g}$	27.8	37.1	3.9	6.7	28.1	37.7	0.0	0.0
Scenario (ii): model for M_T ($\times$)									
Model for POM ($\times$)	$\widehat{\psi}_p$	7.4	20.1	3.5	6.4	8.1	21.1	46.2	17.2
Model for POM ($\checkmark$)	$\widehat{\psi}_{\text{cont},1}$	0.4	0.3	3.4	5.9	3.4	5.9	95.0	95.4
	$\widehat{\psi}_{\text{cont},2}$	0.3	0.3	3.4	5.8	3.4	5.8	95.3	95.6
–	$\widehat{\psi}_{\text{disc},g}$	27.3	39.5	3.9	6.7	27.6	40.0	0.0	0.0

"POM" means the potential outcome mean function $E\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\}$; $\checkmark$ (is correctly specified), $\times$ (is misspecified)

Table S2: Simulation results in Setting II with censoring based on $1,000$ simulated datasets: the Monte Carlo bias, standard error, root mean square error of the estimators, and coverage rate of 95% confidence intervals.

		Bias ($\times 10^3$)		SE ($\times 10^3$)		rMSE ($\times 10^3$)		CR ($\times 10^2$)	
($n = 2000$)	Method	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*	ψ_1^*	ψ_2^*
Scenario (i): model for M_T (✓) and $g \equiv 1$									
Model for POM (×)	$\widehat{\psi}_p$	-2.0	-3.8	46.5	86.3	46.6	86.4	94.5	94.6
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	-1.9	-1.4	45.9	83.1	46.0	83.1	95.4	95.3
–	$\widehat{\psi}_{\text{disc},g}$	273.4	315.2	52.9	94.9	278.4	329.1	1.1	5.9
Scenario (ii): model for M_T (×) and $g \equiv 1$									
Model for POM (×)	$\widehat{\psi}_p$	63.1	175.1	45.5	86.2	77.8	195.2	63.4	35.3
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	4.8	0.5	46.0	83.5	46.3	83.5	94.2	94.4
–	$\widehat{\psi}_{\text{disc},g}$	278.2	319.0	53.3	94.9	283.2	332.8	0.0	2.2
Scenario (iii): model for M_T (✓) and $g = K_C^{-1}$ (✓)									
Model for POM (×)	$\widehat{\psi}_p$	-2.5	-4.1	47.5	87.2	47.6	87.3	94.7	94.5
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	-2.3	-1.3	46.6	83.7	46.7	83.7	95.2	94.9
–	$\widehat{\psi}_{\text{disc},g}$	270.8	310.6	53.6	95.1	276.1	324.8	0.0	2.7
Scenario (iv): model for M_T (×) and $g = K_C^{-1}$ (✓)									
Model for POM (×)	$\widehat{\psi}_p$	63.2	155.8	44.2	84.0	77.1	177.0	57.3	31.7
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	8.8	-4.0	46.7	83.6	47.5	83.7	94.6	95.1
–	$\widehat{\psi}_{\text{disc},g}$	280.9	310.1	54.3	95.8	286.1	324.6	0.0	2.6
Scenario (v): model for M_T (✓) and $g = K_C^{-1}$ (×)									
Model for POM (×)	$\widehat{\psi}_p$	-2.1	-3.8	47.8	88.3	47.8	88.4	95.0	94.8
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	-2.0	-1.0	47.0	84.3	47.0	84.3	95.2	95.4
–	$\widehat{\psi}_{\text{disc},g}$	270.8	310.6	53.6	95.1	276.1	324.8	0.0	1.8
Scenario (vi): model for M_T (×) and $g = K_C^{-1}$ (×)									
Model for POM (×)	$\widehat{\psi}_p$	72.3	189.7	46.5	88.7	86.0	209.4	52.7	29.5
Model for POM (✓)	$\widehat{\psi}_{\text{cont},1}$	12.3	-0.5	46.9	83.8	48.5	83.8	94.6	95.2
–	$\widehat{\psi}_{\text{disc},g}$	280.9	310.1	54.3	95.8	286.1	324.6	0.1	1.1

"POM" means the potential outcome mean function $E\{H(\psi^*) \mid \overline{V}_u, T \geqslant u\}$; ✓ (is correctly specified), × (is misspecified)

Table S3: Results of sensitivity analysis for the effect of time to HAART initiation on the CD4 count at year 2: the unit is cells/mm^3 per month

Parameter	Est	SE	lower .95	upper .95	p-val
ψ_1^*	0.57	11.83	-22.62	23.75	0.96
ψ_2^*	**-0.36**	0.36	-1.07	0.35	0.32
ψ_3^*	3.50	8.08	-12.34	19.33	0.67
ψ_4^*	-0.04	0.19	-0.41	0.33	0.83
ψ_5^*	6.32	4.89	-3.28	15.91	0.20
ψ_6^*	-25.32	14.06	-52.87	2.23	0.07
ψ_7^*	0.40	0.34	-0.27	1.08	0.24
ψ_8^*	0.54	0.66	-0.75	1.83	0.41