BIOMETRICS ***, 1–29 DOI: *******
*** ***

Marginal Proportional Hazards Models for Clustered Interval-Censored Data with Time-Dependent Covariates

Kaitlyn Cook[1,*], **Wenbin Lu**[2,**], **and Rui Wang**[1,3,***]

[1] Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts, U.S.A.

[2] Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A.

[3] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, U.S.A.

**email:* kaitlyn_cook@harvardpilgrim.org

***email:* wlu4@ncsu.edu

****email:* rwang@hsph.harvard.edu

SUMMARY: The Botswana Combination Prevention Project was a cluster-randomized HIV prevention trial whose follow-up period coincided with Botswana's national adoption of a universal test-and-treat strategy for HIV management. Of interest is whether, and to what extent, this change in policy (i) modified the observed preventative effects of the study intervention and (ii) was associated with a reduction in the population-level incidence of HIV in Botswana. To address these questions, we propose a stratified proportional hazards model for clustered interval-censored data with time-dependent covariates and develop a composite expectation maximization algorithm that facilitates estimation of model parameters without placing parametric assumptions on either the baseline hazard functions or the within-cluster dependence structure. We show that the resulting estimators for the regression parameters are consistent and asymptotically normal. We also propose and provide theoretical justification for the use of the profile composite likelihood function to construct a robust sandwich estimator for the variance. We characterize the finite-sample performance and robustness of these estimators through extensive simulation studies. Finally, we conclude by applying this stratified proportional hazards model to a re-analysis of the Botswana Combination Prevention Project, with the national adoption of a universal test-and-treat strategy now modeled as a time-dependent covariate.

KEY WORDS: Clustered failure time data; Composite EM algorithm; Composite likelihood; HIV/AIDS; Interval censoring; Marginal models; Nonparametric likelihood; Proportional hazards; Semiparametric regression; Time-dependent covariates.

1. Introduction

Interval-censored data naturally arise in biomedical and epidemiological studies in which the event of interest is subject to periodic follow-up or otherwise cannot be observed directly: the timing of this event is resolved only up to the interval between successive examinations. When the subjects in these studies also belong to existing, non-investigator determined groups—such as hospitals, communities, social networks, or insurance networks—the resulting data may be both interval-censored and clustered. The conditional analysis of clustered, interval-censored data has recently received increased attention and methodological development (e.g., Zeng *et al.*, 2017; Gao *et al.*, 2019; Yang *et al.*, 2021). Under the conditional modeling framework, within-cluster correlation is explicitly modeled through the inclusion of latent random effects. The parameters of this mixed-effects model are then estimated by maximizing the corresponding full-data likelihood, found by integrating over the distribution of the random effects, and have a cluster-conditional interpretation.

Here our interest lies instead in the marginal analysis of clustered interval-censored data, in which marginal (population-averaged) covariate effects may be estimated without needing to specify the within-cluster correlation structure. To motivate this interest, we introduce the Botswana Combination Prevention Project (BCPP), a cluster-randomized trial (CRT) evaluating whether, and to what extent, holistic combination prevention efforts could reduce the population-level incidence of HIV in Botswana (Makhema *et al.*, 2019). Thirty communities across Botswana were pair-matched and then randomized within pair to either the combination prevention package—which included expanded HIV testing and case identification, enhanced linkage-to-care efforts, an dearly initiation of antiretroviral therapy (ART)—or an enhanced standard of care; HIV incidence was then assessed using a cohort of 8,551 HIV-negative individuals, who were tested on an approximately annual basis for the onset of infection. The resulting time to HIV seroconversion was thus interval-censored and subject to potentially complex within-cluster correlation, with this correlation driven by both the well-documented CRT clustering effect (Hayes and Moulton, 2017) and the unobserved transmission networks within each community. It was also subject to an intercurrent event: in June 2016, while the BCPP was still ongoing, the Botswana Ministry of Health updated the national HIV treatment guidelines to include a universal test-and-treat (UTT) strategy (WHO, 2016). The new policy recommended the immediate initiation of ART for all individuals diagnosed with HIV. As a result, both the standard-of-care and intervention packages provided to individuals in the BCPP communities changed over the course of follow-up. These changes raise several questions about the effect of combination HIV prevention in the BCPP communities and about HIV management in Botswana more generally. In particular: (i) What implications, if any, did the mid-trial changes to the standard-of-care and intervention packages have for the final estimated benefit of combination HIV prevention? (ii) Was the national adoption of a UTT strategy itself associated with a significant reduction in HIV incidence in Botswana?

Answering these questions requires contending with several methodological challenges. For right-censored failure times, methods based on the theory of generalized estimating equations have been developed to estimate marginal covariate effects for potentially time-dependent covariates without needing to assume any particular underlying correlation structure (Wei *et al.*, 1989; Cai *et al.*, 2000; Lin, 1994). But extending these methods to interval-censored settings faces additional challenges. As a consequence of the interval censoring mechanism, we observe only a partial ordering of the underlying failure times and their accompanying risk sets, which in turn precludes construction of the familiar partial likelihood for semi-parametric proportional hazards models. For independent interval-censored data, Satten (1996) and Goggins *et al.* (1998) consider the distribution of possible rankings of failure times that are consistent with the observed censoring intervals and thereby circumvent estimation of the baseline hazard function. Other estimating approaches generally involve estimation of the baseline hazard function(s) – with the dimensionality of the corresponding nonparametric estimator(s) growing in lockstep with the number of unique monitoring times in the sample. Setting aside the theoretical challenges that this nonparametric estimation poses, it also has practical consequences for the use of generalized estimating equations to fit marginal proportional hazards models to correlated interval-censored data: each update of the baseline hazard parameters is computationally intensive, and both the Fisher scoring and variance estimation procedures now require the inversion of high-dimensional matrices. Existing marginal methods thus seek to bound the dimensionality of the baseline hazard parameters, either by assuming a parametric form for the baseline survival distribution (Cook and Tolusso, 2009; Zhang and Sun, 2010), placing a restriction on the number of permissible monitoring times (Goggins and Finkelstein, 2000; Kim and Xue, 2002; Tong *et al.*, 2008; Chen *et al.*, 2013), or approximating the baseline hazard with a piecewise-constant function (Zhang and Sun, 2013; Kor *et al.*, 2013). To the best of our knowledge, no unified framework exists for the marginal analysis of clustered interval-censored data under the proportional hazards framework with nonparametric estimation of the baseline hazard function(s).

Here we aim to redress this gap. We draw on the theory of composite likelihood functions in order to (i) construct a set of unbiased estimating equations for the model parameters and (ii) develop a corresponding composite expectation maximization algorithm; neither requires specification of any higher order moments of the observed-data distribution. Our optimization procedure readily incorporates both time-independent and time-dependent covariates as well as fully nonparametric estimation of the baseline hazard function(s), which we allow to be either stratified or unstratified. It also results in closed-form updating equations for the baseline hazard estimators, so that the estimation procedure easily scales settings such as the BCPP in which the number of unique monitoring times in the sample may be large.

We show that the resulting estimators for the regression parameters are consistent and asymptotically normal. To facilitate interval estimation and inference, we develop a profile composite likelihood-based variance estimator that is similarly robust to the within-cluster dependence structure. We motivate this estimator by deriving the quadratic expansion of the profile composite log-likelihood function, which extends the results of Murphy and van der Vaart (2000)—who demonstrate that the profile likelihood forms a valid basis for inference with semiparametric efficient estimators—to the independence composite likelihood setting.

Section 2 introduces our setting and notation, and presents the marginal proportional hazards model of interest. Section 3 discusses estimation and inference for this model: we introduce the robust estimating function in Section 3.1, propose an optimization algorithm for this function in Section 3.2, and both (i) demonstrate the strong consistency and weak convergence of the resulting estimator and (ii) develop our robust variance estimation approach in Section 3.3. We study the finite sample properties of the resulting point and interval estimators through a series of simulation studies in Section 4, before returning to our motivating analysis of UTT adoption in Botswana in Section 5. Section 6 concludes with a brief summary and discussion.

2. Setting and Notation

Consider the setting in which the data comprise M clusters and let n_i represent the number of subjects within cluster i. Suppose as well that the outcome of interest is the time to some event, such as the onset of infection, and let T_{ij} denote this event time for subject j in cluster i. We assume that the corresponding marginal hazard is given by

$$\lambda_{ij}\{t|\boldsymbol{X}_{ij}(t), \boldsymbol{Z}_{ij}\} = \lambda_{\boldsymbol{Z}_{ij}}(t)\exp\{\boldsymbol{\beta}^{\top}\boldsymbol{X}_{ij}(t)\}, \tag{1}$$

where $\boldsymbol{Z}_{ij}$ is a discrete random vector containing the set of stratification factors, which we assume takes on $S \in \mathbb{N}$ distinct levels denoted by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_S$; $\lambda_{\boldsymbol{z}_s}(t) \equiv \lambda_s(t)$ is an arbitrary stratum-specific baseline hazard function, shared among all observations with $\boldsymbol{Z}_{ij} = \boldsymbol{z}_s$

for $s = 1, \dots, S$; $\boldsymbol{X}_{ij}(t)$ is a p-dimensional bounded covariate process informing the time-to-event distribution; and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a p-dimensional vector of covariate effects. Note that model (1) permits great flexibility in modeling the event of interest: it reduces to the traditional stratified proportional hazards model in the event that $\boldsymbol{X}_{ij}(t)$ is time-invariant, $\boldsymbol{X}_{ij}(t) \equiv \boldsymbol{X}_{ij}$, and to the traditional unstratified proportional hazards model in the event that $\lambda_s(t) = \lambda(t)$ for all levels $\boldsymbol{z}_s$ of the stratification factor. It also readily incorporates additive time-varying covariate effects, $\beta(t) = \beta_0 + \sum_{l=1}^m \beta_l f_l(t)$ with $f_1, \dots, f_m$ being known functions of time, as $\beta(t)X_{ij} \equiv \boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)$ under $\boldsymbol{\beta} = (\beta_0, \dots, \beta_m)^\top$ and $\boldsymbol{X}_{ij}(t) = (X_{ij}, f_1(t)X_{ij}, \dots, f_m(t)X_{ij})^\top$.

Suppose further that the occurrence of the event of interest is monitored only through a series of periodic examinations or tests. Let K_{ij} be a positive, integer-valued random variable indicating the number of examinations on subject j in cluster i, and let $\boldsymbol{Y}_{ij} = \{Y_{ijk} : k = 1, \dots, K_{ij}\}$ be a vector of random length containing the corresponding sequence of monitoring times, $Y_{ij1} < \cdots < Y_{ijK_{ij}}$. Then T_{ij} is known only up to the interval $(L_{ij}, U_{ij}]$, where $(L_{ij}, U_{ij}]$ is defined as the interval among $[0, Y_{ij1}], (Y_{ij1}, Y_{ij2}], \dots, (Y_{ijK_{ij}}, \infty)$ that contains the true T_{ij}. We assume that this monitoring process is non-informative, i.e., that

$$\{K_{ij}, \boldsymbol{Y}_{ij} : j = 1, \dots, n_i\} \perp\!\!\!\perp \{T_{ij} : j = 1, \dots, n_i\} | \{\boldsymbol{Z}_{ij}, \boldsymbol{X}_{ij}(\cdot) : j = 1, \dots, n_i\},$$

and that the resulting interval-censoring occurs only to the outcome, i.e., that $(\boldsymbol{X}_{ij}(t) : t \leqslant U^*_{ij})$ is otherwise fully observed, with $U^*_{ij} := L_{ij}I(U_{ij} = \infty) + U_{ij}I(U_{ij} < \infty)$ denoting the total time that subject j in cluster i is under monitoring for the event of interest.

Finally, let $\mathcal{O}_{ij} = \{L_{ij}, U_{ij}, \boldsymbol{Z}_{ij}, (\boldsymbol{X}_{ij}(t) : t \leqslant U^*_{ij})\}$ be the full collection of observed data for subject j in cluster i. We assume that these collections are independent across clusters, so that $\mathcal{O}_{ij} \perp\!\!\!\perp \mathcal{O}_{i'j'}$ for all $i \neq i'$, but make no further assumptions about the within-cluster dependence structure. Our aim is to then use these correlated, interval-censored data to conduct semiparametric estimation and inference for model (1).

3. Methods

Section 3.1 provides a brief overview of composite likelihood theory, which we use to construct an estimating function for (1) that is agnostic to the higher-order structure of the observed data. We maximize this function using a variant of the classical expectation maximization algorithm that has been adapted to the composite likelihood setting in Section 3.2. Section 3.3 develops asymptotic theory for the resulting maximum composite likelihood estimators and proposes a robust variance estimator for the parametric component.

3.1 *Independence Composite Likelihood for (1) Under Nonparametric Estimation of the Baseline Hazard*

The composite likelihood paradigm facilitates inference in high-dimensional or correlated-data settings through what is effectively an act of dimension reduction: it constructs a full-data pseudo-likelihood by multiplying together a collection of lower-dimensional component densities (Varin *et al.*, 2011). For example, the independence composite likelihood function is found by multiplying together the univariate marginal density functions:

$$\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}}) = \prod_{i=1}^{M} \prod_{j=1}^{n_i} f(\mathcal{O}_{ij}; \theta). \tag{2}$$

As each component of (2) is a valid density, the resulting composite score equation, $u(\theta) = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \partial \log f(\mathcal{O}_{ij}; \theta)/\partial\theta$, forms an unbiased estimating equation; the solution is called the maximum composite likelihood estimator, $\widehat{\theta}_{CL}$. $\widehat{\theta}_{CL}$ has been shown in parametric settings to be consistent and asymptotically normal, and is robust in the sense that it provides valid inference for the class of full-data joint distributions that are consistent with the component densities in $\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$ (Chandler and Bate, 2007; Varin *et al.*, 2011; Xu and Reid, 2011). In this way, the composite likelihood function allows one to conduct estimation and inference for marginal parameters of the full-data distribution *without* needing to specify any of its higher-order moments; it is thus ideally suited to the analysis of infectious disease prevention studies, cluster-randomized trials, and other settings in which complex dependencies may exist between the observations.

Let $\theta = (\boldsymbol{\beta}, \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda} = (\Lambda_1, \ldots, \Lambda_S)$ and $\Lambda_s(\cdot) = \int_0^{(\cdot)} \lambda_s(t)dt$. Then the independence composite likelihood for θ is

$$\begin{aligned}\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}}) &= \prod_{i=1}^{M}\prod_{j=1}^{n_i}\left[\exp\left\{-\int_0^{L_{ij}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_{\boldsymbol{Z}_{ij}}(t)\right\} - \left\{-\int_0^{U_{ij}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_{\boldsymbol{Z}_{ij}}(t)\right\}\right] \qquad (3)\\ &= \prod_{i=1}^{M}\prod_{j=1}^{n_i}\prod_{s=1}^{S}\left[\exp\left\{-\int_0^{L_{ij}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t)\right\} - \left\{-\int_0^{U_{ij}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t)\right\}\right]^{I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)}.\end{aligned}$$

Note that, under the independence composite likelihood construction, the terms in (3) may be reindexed and rearranged without changing $\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$. To facilitate estimation of the baseline hazard functions $\Lambda_1, \ldots, \Lambda_S$ and to simplify notation, it will be advantageous to reindex the product in $\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$ so that the data are partitioned according to strata defined by levels of $\boldsymbol{Z}_{ij}$ as opposed to clusters indexed by i; we will use this reindexing throughout Section 3.1 and Section 3.2, but not for the asymptotic results presented in Section 3.3. In particular, let s ($s = 1, \ldots, S$) index the strata and v ($v = 1, \ldots, n_s$) index the subjects within each stratum. Then (3) is equivalently given by

$$\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}}) = \prod_{s=1}^{S}\prod_{v=1}^{n_s}\left[\exp\left\{-\int_0^{L_{sv}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{sv}(t)} d\Lambda_s(t)\right\} - \left\{-\int_0^{U_{sv}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{sv}(t)} d\Lambda_s(t)\right\}\right].$$

Finally, to present $\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$ under nonparametric estimation of $\Lambda_1, \ldots, \Lambda_S$, we establish some additional notation regarding the stratum-specific baseline hazard functions. Let $\boldsymbol{\tau}_s$ be the ordered set of all unique $L_{sv} > 0$ and $U_{sv} < \infty$ in stratum s, and take $|\boldsymbol{\tau}_s| = \rho_s$. The nonparametric maximum composite likelihood estimator for Λ_s is restricted to the class of non-decreasing step functions with potential discontinuities at times $\tau_{sr} \in \boldsymbol{\tau}_s$ only; we parametrize the step size at τ_{sr} by $\lambda_{sr} \geqslant 0$ so that $\Lambda_s(t) := \sum_{\tau_{sr} \leqslant t} \lambda_{sr}$. The independence composite likelihood for $\theta = (\boldsymbol{\beta}, \lambda_{11}, \ldots, \lambda_{S\rho_S})^\top$ is then

$$\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}}) = \prod_{s=1}^{S}\prod_{v=1}^{n_s} \exp\left(-\sum_{\tau_{sr}\leqslant L_{sv}} \lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{svr}}\right)\left\{1-\exp\left(-\sum_{L_{sv}<\tau_{sr}\leqslant U_{sv}} \lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{svr}}\right)\right\}^{I(U_{sv}<\infty)} \tag{4}$$

with $\boldsymbol{X}_{svr} \equiv \boldsymbol{X}_{sv}(\tau_{sr})$.

3.2 *Expectation Maximization for Composite Likelihoods*

Direct maximization of (4) to obtain $\widehat{\theta}_{CL}$ is challenging. This is in large part due to the interval censoring mechanism, which prevents observation of either the exact failure times or their associated counting process. In other words, the observed interval-censored data, $\boldsymbol{\mathcal{O}}$, represent a many-to-one function (or coarsening) of an augmented data collection, $\boldsymbol{\mathcal{A}}$, for which estimation of θ would be more straightforward. In traditional maximum likelihood settings, the expectation maximization (EM) algorithm recasts maximization of $\log \mathcal{L}(\theta; \boldsymbol{\mathcal{O}}) \equiv \log f(\boldsymbol{\mathcal{O}}; \theta)$ in terms of these augmented data by iteratively maximizing the closest approximation to $\log \mathcal{L}(\theta; \boldsymbol{\mathcal{A}}) \equiv \log f(\boldsymbol{\mathcal{A}}; \theta)$ given the observed data: its projection onto the observed data model. The algorithm thus alternates between an expectation step, which constructs the objective function $\mathcal{Q}(\theta|\theta_l) = \mathbb{E}_{\theta_l}[\log \mathcal{L}(\theta; \boldsymbol{\mathcal{A}})|\boldsymbol{\mathcal{O}}]$ at the current parameter estimate θ_l, and a maximization step, which updates $\theta_{l+1} = \text{argmax}\mathcal{Q}(\theta|\theta_l)$. Provided that it exists and is unique, $\widehat{\theta}_{MLE} = \text{argmax} \log \mathcal{L}(\theta; \boldsymbol{\mathcal{O}})$ is a fixed point solution of the algorithm.

Here, however, we wish to *avoid* specification of the higher-order moments of $f(\boldsymbol{\mathcal{O}}; \theta)$, $f(\boldsymbol{\mathcal{A}}; \theta)$, and $f(\boldsymbol{\mathcal{A}}|\boldsymbol{\mathcal{O}}; \theta)$ when maximizing $\log \mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$, so as to maintain the robustness property of $\widehat{\theta}_{CL}$. To that end, Gao and Song (2011) propose a composite likelihood analog of the EM algorithm, in which the observed, augmented, and conditional data distributions are each replaced by a working composite "model" (formed, as in the composite likelihood setting, by multiplying together a collection of component densities), and in which the maximum *composite* likelihood estimator is a fixed point solution of the procedure. It requires only that the component densities of these working models are congenial with one another, and makes no assumptions about their relationship to the true joint distributions. See Web Appendix A.1 for further discussion of the composite EM algorithm and its properties.

We now present a composite EM algorithm for maximization of the independence composite likelihood in (4). Following the lead of Zeng *et al.* (2017), we introduce for each subject v in stratum s a collection of ρ_s independent Poisson random variables: $W_{svr} \sim Pois\{\lambda_{sr}\exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{svr})\}$ for $r = 1, \ldots, \rho_s$. We also define the random variables $A_{sv} := \sum_{\tau_{sr} \leqslant L_{sv}} W_{svr}$ and $B_{sv} := I(U_{sv} < \infty) \sum_{L_{sv} < \tau_{sr} \leqslant U_{sv}} W_{svr}$.

Remark 1: To gain intuition for these augmentation variables, let $N_{sv}(t)$ be a Poisson counting process with intensity function (1). Then $\boldsymbol{W}_{sv} = \{W_{svr} : r = 1, \ldots, \rho_s\}$ can be understood as a sequence of independent increments from this process, with W_{svr} counting the number of events in the interval $(\tau_{s,r-1}, \tau_{sr}]$. Thus $A_{sv} = \sum_{\tau_{sr} \leqslant L_{sv}} W_{svr} \equiv N_{sv}(L_{sv})$ corresponds to the number of events experienced by subject v in stratum s prior to time L_{sv}, and $B_{sv} = I(U_{sv} < \infty) \sum_{L_{sv} < \tau_{sr} \leqslant U_{sv}} W_{svr} \equiv N_{sv}(U^*_{sv}) - N_{sv}(L_{sv})$ to the number of events between times L_{sv} and U^*_{sv}. The censoring interval $(L_{sv}, U_{sv}]$ is then equivalent to the event $\{A_{sv} = 0\} \cap \{B_{sv} = 0\}$ for $U_{sv} = \infty$ and the event $\{A_{sv} = 0\} \cap \{B_{sv} > 0\}$ for $U_{sv} < \infty$, so that the observed censoring interval represents a coarsening of the augmented outcome process $\boldsymbol{W}_{sv}$. A formal exposition of this relationship is provided in Web Appendix A.2.

We define the individual observed and Poisson-augmented data collections by $\mathcal{O}_{sv} = \{L_{sv}, U_{sv}, \boldsymbol{Z}_{sv}, (\boldsymbol{X}_{sv}(t) : t \leqslant U^*_{sv})\}$ and $\mathcal{A}_{sv} = \{L_{sv}, U_{sv}, \boldsymbol{W}_{sv}, \boldsymbol{Z}_{sv}, (\boldsymbol{X}_{sv}(t) : t \leqslant U^*_{sv})\}$, respectively, and construct a composite EM algorithm by iteratively (i) taking the projection of the independence composite augmented-data log-likelihood onto the independence model for the observed data and (ii) maximizing the resulting conditional expectation.

Towards this first (expectation) step, we note that the independence composite likelihood for the augmented data is given by

$$\mathcal{L}_C(\theta; \boldsymbol{\mathcal{A}}) = \prod_{s=1}^{S} \prod_{v=1}^{n_s} \prod_{r=1}^{\rho_s} \left\{ \frac{1}{W_{svr}!} \left(\lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{svr}} \right)^{W_{svr}} \exp\left(-\lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{svr}} \right) \right\}^{I(\tau_{sr} \leqslant U^*_{sv})},$$

so that the corresponding composite log-likelihood is, up to a constant,

$$\log \mathcal{L}_C(\theta; \boldsymbol{\mathcal{A}}) \doteq \sum_{s=1}^{S} \sum_{v=1}^{n_s} \sum_{r=1}^{\rho_i} I(\tau_{sr} \leqslant U^*_{sv}) \left\{ W_{svr} \log \left(\lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{svr}} \right) - \lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{svr}} \right\}.$$

Letting θ_l be the current estimate of θ, the composite EM objective function is then simply

$$\mathcal{Q}_C(\theta|\theta_l) = \sum_{s=1}^{S} \sum_{v=1}^{n_s} \sum_{r=1}^{\rho_i} I(\tau_{sr} \leqslant U^*_{sv}) \left\{ \mathbb{E}_{\theta_l}[W_{svr}|\mathcal{O}_{sv}] \log \left(\lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{svr}} \right) - \lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{svr}} \right\}, \quad (5)$$

where

$$\mathbb{E}_{\theta_l}[W_{svr}|\mathcal{O}_{sv}] = \begin{cases} \frac{\lambda_{l,sr} \exp\left(\boldsymbol{\beta}_l^\top \boldsymbol{X}_{svr}\right)}{1 - \exp\left\{ - \sum_{L_{sv} < \tau_{sr} \leqslant U_{sv}} \lambda_{l,sr} \exp\left(\boldsymbol{\beta}_l^\top \boldsymbol{X}_{svr}\right)\right\}} & \text{for } L_{sv} < \tau_{sr} \leqslant U^*_{sv} \\ 0 & \text{otherwise} \end{cases}.$$

Details for the derivation of this conditional expectation are available in Web Appendix A.3. To simplify notation, we use $\widehat{W}_{svr}$ for $\mathbb{E}_{\theta_l}[W_{svr}|\mathcal{O}_{sv}]$ throughout the remainder of Section 3.

The second (maximization) step updates the estimates of θ using a profile likelihood approach. To that end, we first differentiate $\mathcal{Q}_C(\theta|\theta_l)$ with respect to λ_{sr} $(s = 1, \ldots, S; r = 1, \ldots, \rho_s)$ to obtain the following formula:

$$\widehat{\lambda}_{sr}(\boldsymbol{\beta}) = \frac{\sum_{v=1}^{n_s} I(\tau_{sr} \leqslant U^*_{sv}) \widehat{W}_{svr}}{\sum_{v=1}^{n_s} I(\tau_{sr} \leqslant U^*_{sv}) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{svr})}. \quad (6)$$

The profile composite objective function for $\boldsymbol{\beta}$ is then obtained by substituting (6) into (5),

$$\mathcal{Q}_C(\boldsymbol{\beta}, \widehat{\lambda}_{11}(\boldsymbol{\beta}), \ldots, \widehat{\lambda}_{S\rho_S}(\boldsymbol{\beta})|\theta_l) =$$

$$\sum_{s=1}^{S} \sum_{v=1}^{n_s} \sum_{r=1}^{\rho_i} I(\tau_{sr} \leqslant U^*_{sv}) \left\{ \widehat{W}_{svr} \log \left(\frac{\sum_{v'=1}^{n_s} I(\tau_{sr} \leqslant U^*_{sv'}) \widehat{W}_{sv'r}}{\sum_{v'=1}^{n_s} I(\tau_{sr} \leqslant U^*_{sv'}) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{sv'r})} \right) + \widehat{W}_{svr} \boldsymbol{\beta}^\top \boldsymbol{X}_{svr} - \widehat{W}_{svr} \right\},$$

and we update $\boldsymbol{\beta}$ by solving the corresponding score equation:

$$\sum_{s=1}^{S} \sum_{v=1}^{n_s} \sum_{r=1}^{\rho_s} I(\tau_{sr} \leqslant U^*_{sv}) \widehat{W}_{svr} \left\{ \boldsymbol{X}_{svr} - \frac{\sum_{v'=1}^{n_s} I(\tau_{sr} \leqslant U^*_{sv'}) \boldsymbol{X}_{sv'r} \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{sv'r})}{\sum_{v'=1}^{n_s} (\tau_{sr} \leqslant U^*_{sv'}) \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{sv'r})} \right\} \overset{set}{=} 0.$$

Let $\boldsymbol{\beta}_{l+1}$ be the resulting estimate of $\boldsymbol{\beta}$. We update the remaining parameter estimates by setting $\lambda_{l+1,sr} = \widehat{\lambda}_{sr}(\boldsymbol{\beta}_{l+1})$ $(s = 1, \ldots, S; r = 1, \ldots, \rho_s)$. The composite EM algorithm iterates between these expectation and maximization steps until convergence, here defined to be $||\boldsymbol{\beta}_{l+1} - \boldsymbol{\beta}_l||_1 < \epsilon_{tol}$. We denote the final estimators by $\widehat{\theta}_{CL} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$, with $\widehat{\boldsymbol{\Lambda}} = (\widehat{\Lambda}_1, \ldots, \widehat{\Lambda}_S)$.

3.3 *Asymptotic Theory*

We establish the asymptotic properties of $\widehat{\theta}_{CL}$ under the setting in which the number of clusters $M \to \infty$. To do so, we first introduce some additional notation regarding the first and second partial derivatives of the independence composite log-likelihood function. Let $m(\boldsymbol{\beta}, \boldsymbol{\Lambda}; \boldsymbol{\mathcal{O}}_i) = \log\left\{\prod_{j=1}^{n_i} f(\mathcal{O}_{ij}; \theta)\right\}$ denote the individual cluster-level contribution to the log of the independence composite likelihood function in (3) and take $\boldsymbol{\Lambda}_{\epsilon,\boldsymbol{h}}$ to be the parametric submodel for $\boldsymbol{\Lambda}$ satisfying the relationship $d\boldsymbol{\Lambda}_{\epsilon,\boldsymbol{h}} = ((1+\epsilon h_1)d\Lambda_1, \ldots, (1+\epsilon h_S)d\Lambda_S)^\top$ for $\epsilon \in \mathbb{R}$, $\boldsymbol{h} = (h_1, \ldots, h_S)$, and $h_s \in L_2(\mu)$; we take μ to be a dominating measure on the support of the monitoring times, $\mathcal{Y}$. Then

$$m_1(\boldsymbol{\beta}, \boldsymbol{\Lambda}) := \frac{\partial}{\partial \boldsymbol{\beta}^\top} m(\boldsymbol{\beta}, \boldsymbol{\Lambda}; \boldsymbol{\mathcal{O}}_i) \qquad m_2(\boldsymbol{\beta}, \boldsymbol{\Lambda})[\boldsymbol{h}] := \frac{\partial}{\partial \epsilon} m(\boldsymbol{\beta}, \boldsymbol{\Lambda}_{\epsilon,\boldsymbol{h}}; \boldsymbol{\mathcal{O}}_i)\Big|_{\epsilon=0}$$

are the independence composite score equation for $\boldsymbol{\beta}$ and the independence composite score operator for $\boldsymbol{\Lambda}$, respectively, and we define

$$m_{11}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) := \frac{\partial}{\partial \boldsymbol{\beta}} m_1(\boldsymbol{\beta}, \boldsymbol{\Lambda}) \qquad m_{21}(\boldsymbol{\beta}, \boldsymbol{\Lambda})[\boldsymbol{h}] := \frac{\partial}{\partial \boldsymbol{\beta}} m_2(\boldsymbol{\beta}, \boldsymbol{\Lambda})[\boldsymbol{h}].$$

The exact form of m_1, m_2, m_{11}, and m_{21} are given in equations (S.7)–(S.10) of Web Appendix B. Then under the regularity conditions given in Web Appendix B, we may establish the strong consistency of $\widehat{\theta}_{CL}$ (Theorem 1) and the asymptotic distribution of its parametric component, $\widehat{\boldsymbol{\beta}}$ (Theorem 2):

Theorem 1: *Under Conditions 1–6,* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \sum_{s=1}^{S} \|\widehat{\Lambda}_s - \widehat{\Lambda}_{s0}\|_{l^\infty(\mathcal{Y})} \overset{a.s.}{\to} 0$*, where* $\|\cdot\|$ *is the standard Euclidean norm and* $\|\cdot\|_{l^\infty(\mathcal{Y})}$ *is the supremum norm on* $\mathcal{Y}$*.*

Theorem 2: *Under Conditions 1–8,* $\sqrt{M}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ *converges weakly to a p-dimensional zero-mean normal random vector with covariance matrix*

$$\begin{aligned} I^* = {} & \mathbb{E}_{\theta_0}\left\{m_{11}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_{21}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}^{-1} \\ & \times \mathbb{E}_{\theta_0}\left[\left\{m_1(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_2(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}^{\otimes 2}\right] \\ & \times \mathbb{E}_{\theta_0}\left\{m_{11}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_{21}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}^{-1}, \end{aligned} \tag{7}$$

where $\boldsymbol{h}^* = (h_1^*, \ldots, h_S^*)$*,* $h_s^* \in \mathcal{L}_2(\mu)$*, satisfies equation (S.13) in Web Appendix B.*

Remark 2: The proofs of all theorems are deferred to Web Appendix B. Briefly, the proof of Theorem 1 conceives of the maximum composite likelihood estimator as a semiparametric M estimator with criterion function $m(\boldsymbol{\beta}, \boldsymbol{\Lambda}; \boldsymbol{\mathcal{O}}_i)$. We establish that $(\boldsymbol{\beta}_0, \Lambda_{10}, \ldots, \Lambda_{S0})$ is a unique and well-separated maximizer of the population version of this criterion function and then apply the Argmax Theorem for semiparametric M estimators (Theorem 2.12 of Kosorok, 2008) to conclude almost sure convergence of $(\widehat{\boldsymbol{\beta}}, \widehat{\Lambda}_1, \ldots, \widehat{\Lambda}_S)$ to this maximizer. In the proof of Theorem 2, we use techniques from empirical process theory to establish the $M^{1/3}$ convergence of $\widehat{\boldsymbol{\Lambda}}$ and the stochastic equicontinuity of function classes related to m_1 and m_2; we then apply Taylor expansions to the independence composite score function and independence composite score operator to arrive at the limiting distribution of $\sqrt{M}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$.

To arrive at an estimator for the covariance matrix of $\widehat{\boldsymbol{\beta}}$, we introduce the *profile composite log-likelihood function* for $\boldsymbol{\beta}$:

$$p\ell_C(\boldsymbol{\beta}) := \sup_{\boldsymbol{\Lambda} \in \mathcal{C}} \log \mathcal{L}_C(\boldsymbol{\beta}, \boldsymbol{\Lambda}; \boldsymbol{\mathcal{O}}) = \log \mathcal{L}_C \left\{ \boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta}); \boldsymbol{\mathcal{O}} \right\}, \tag{8}$$

where $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta}) = \operatorname{argmax}_{\boldsymbol{\Lambda} \in \mathcal{C}} \log \mathcal{L}_C(\boldsymbol{\beta}, \boldsymbol{\Lambda}; \boldsymbol{\mathcal{O}})$ and $\mathcal{C} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_S$ with $\mathcal{C}_s$ the set of step functions with non-negative jumps at times τ_{sr} $(r = 1, \ldots, \rho_s)$. Theorem 3 provides the second-order Taylor expansion of this profile composite log-likelihood function about $\boldsymbol{\beta}_0$.

Theorem 3: *Under Conditions 1–8, for any sequence* $\widetilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$*,*

$$\begin{aligned} p\ell_C(\widetilde{\boldsymbol{\beta}}) &= p\ell_C(\boldsymbol{\beta}_0) \\ &\quad + (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \sum_{i=1}^{M} \left\{ m_1(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_2(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*] \right\} (\boldsymbol{\mathcal{O}}_i) \qquad (9) \\ &\quad - \frac{1}{2} M (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbb{E}_{\theta_0} \left\{ m_{11}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_{21}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*] \right\} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_P(1 + \sqrt{M} \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|)^2. \end{aligned}$$

Remark 3: Murphy and van der Vaart (2000) prove a similar asymptotic expansion for the profile log-likelihood function under semiparametric maximum likelihood estimation and then use this quadratic form to justify the use of the profile likelihood function as though an ordinary likelihood when conducting inference on low-dimensional parameters (cf. Theorem 1 of Murphy and van der Vaart (2000)). Theorem 3 extends their profile likelihood expansion to the independence composite likelihood setting.

The asymptotic expansion in (9) suggests that I^* is the inverse of the Godambe information about $\boldsymbol{\beta}$ in a single cluster under the profile composite likelihood function, a result that parallels the form of the asymptotic covariance matrix for parametric maximum composite likelihood estimators (see Varin *et al.*, 2011). It thus provides an immediate path forward for estimating $Cov(\widehat{\boldsymbol{\beta}})$: the first and third terms in (7) may be estimated by the curvature of $p\ell_C(\boldsymbol{\beta})$ at $\widehat{\boldsymbol{\beta}}$ and the second term by the variance of its gradient. To formalize this observation, let $H(\boldsymbol{\beta}) := \mathbb{E}_{\theta_0}\left\{m_{11}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_{21}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}$ and $J(\boldsymbol{\beta}) := \mathbb{E}_{\theta_0}\left[\{m_1(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_2(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\}^{\otimes 2}\right]$ be the sensitivity and variability matrices, respectively; let $\boldsymbol{e}_k$ be the kth canonical vector in $\mathbb{R}^p$; and let h_M be a perturbation constant of order $M^{-1/2}$. Then we approximate the (k,l)th entry of the sample sensitivity matrix by

$$\left[\widehat{H}_M(\widehat{\boldsymbol{\beta}})\right]_{kl} = \frac{p\ell_C(\widehat{\boldsymbol{\beta}}; \boldsymbol{\mathcal{O}}) - p\ell_C(\widehat{\boldsymbol{\beta}} + h_M\boldsymbol{e}_k; \boldsymbol{\mathcal{O}}) - p\ell_C(\widehat{\boldsymbol{\beta}} + h_M\boldsymbol{e}_l; \boldsymbol{\mathcal{O}}) + p\ell_C(\widehat{\boldsymbol{\beta}} + h_M\boldsymbol{e}_k + h_M\boldsymbol{e}_l; \boldsymbol{\mathcal{O}})}{h_M^2}$$

and the sample variability matrix by

$$\widehat{J}_M(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{M} \begin{pmatrix} \{p\ell_C(\widehat{\boldsymbol{\beta}} + h_M\boldsymbol{e}_1; \mathcal{O}_i) - p\ell_C(\widehat{\boldsymbol{\beta}}; \mathcal{O}_i)\}/h_M \\ \vdots \\ \{p\ell_C(\widehat{\boldsymbol{\beta}} + h_M\boldsymbol{e}_p; \mathcal{O}_i) - p\ell_C(\widehat{\boldsymbol{\beta}}; \mathcal{O}_i)\}/h_M \end{pmatrix}^{\otimes 2},$$

where $p\ell_C(\boldsymbol{\beta}; \mathcal{O}_i)$ is the cluster-level contribution to the profile composite log-likelihood function, $p\ell_C(\boldsymbol{\beta}; \mathcal{O}_i) = \log \mathcal{L}_C(\boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta}); \mathcal{O}_i)$. Then

$$\widehat{Cov}(\widehat{\boldsymbol{\beta}}) = M^{-1}\widehat{I}^* = \widehat{H}_M(\widehat{\boldsymbol{\beta}})^{-1}\widehat{J}_M(\widehat{\boldsymbol{\beta}})\widehat{H}_M(\widehat{\boldsymbol{\beta}})^{-1}.$$

In order to implement this robust variance estimator, we require a means to evaluate the profile composite log-likelihood in (8)—or, equivalently, to obtain the profile composite log-likelihood maximizers, $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta})$—at chosen values of $\boldsymbol{\beta}$. We do so by implementing the composite

EM algorithm detailed in Section 3.2, though now with $\boldsymbol{\beta}$ fixed at its chosen value throughout and with only the λ_{sr} $(s = 1, \ldots, S; r = 1, \ldots, \rho_s)$ updated during the maximization step. The algorithm iterates until $\| \log \mathcal{L}_C\{\boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}_{l+1}(\boldsymbol{\beta})\} - \log \mathcal{L}_C\{\boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}_l(\boldsymbol{\beta})\}\|_1 < \epsilon_{\text{tol}}$.

4. Finite-Sample Performance

We conducted a series of simulation studies to assess the finite-sample performance of our proposed point and interval estimators under a range of clustered data structures, marginal hazard specifications, and monitoring schedules.

4.1 *Primary Simulation Study*

We simulated data under two scenarios for the clustering and stratification factors:

I. We assumed that the data comprised a large number of small clusters, with $M \in \{100, 200\}$ and $n_i \sim Unif(20, 30)$. We took the stratification factor $Z_{ij} \sim Unif\{1, 2, 3, 4\}$, so that Z_{ij} was cluster-varying and the number of strata $S = 4$.

II. Drawing on the motivating pragmatic CRT setting, in which only a small number of large clusters were randomized, we assumed that the data comprised $M \in \{15, 20\}$ matched pairs of communities, with community p $(p = 1, 2)$ in matched pair i comprising $n_i^p \sim Unif(250, 350)$ subjects; then $500 \leqslant n_i \leqslant 700$. We also considered one setting in which $M = 15$, $n_i^p \sim Unif(400, 500)$, and $800 \leqslant n_i \leqslant 1000$. Finally, we took the stratification factor Z_{ij} to be an indicator of matched pair membership, so that Z_{ij} was cluster constant and the number of strata $S = M$.

For each scenario, we considered two within-cluster correlation structures: an exchangeable structure, in which subjects were independent conditional on cluster membership, and a hierarchical structure, in which subjects within each cluster were further grouped into independent sub-clusters (representing, for example, families or sexual partnerships) whose sizes varied according to a $Pois(2)$ distribution with a minimum cluster size of 2; dependence within each sub-cluster was modeled using a Clayton copula with Kendall's $\tau = 0.5$.

Each subject j in cluster i was monitored for an event at $K_{ij} = 20$ time points, with the gap times $Y_{ij,k+1} - Y_{ijk} \sim Unif(0, 16)$ and the maximum duration of follow-up given by $\Upsilon = \max_{ij} Y_{ij20}$. We considered the following three models for the time to event:

$$\lambda_{ij}(t|X_{ij}, Z_{ij}) = \lambda_{Z_{ij}}(t) \exp(\beta X_{ij}) \tag{10}$$

$$\lambda_{ij}(t|X_{ij}, Z_{ij}) = \lambda_{Z_{ij}}(t) \exp\{(\gamma_1 + \gamma_2 \log t)X_{ij}\} \tag{11}$$

$$\lambda_{ij}\{t|X_{ij}, P_{ij}(t), Z_{ij}\} = \lambda_{Z_{ij}}(t) \exp\{\alpha_1 X_{ij} + \alpha_2 P_{ij}(t) + \alpha_3 X_{ij} P_{ij}(t)\} \tag{12}$$

where X_{ij} was a binary covariate that was constant within clusters (scenario I) or within communities (scenario II) and $P_{ij}(t) = I(t \geqslant \upsilon_{ij})$ was a time-dependent indicator variable with $\upsilon_{ij} \sim Unif(0, \Upsilon)$. We generated the stratum-specific baseline hazard functions according to the random spline method of Harden and Kropko (2019) in scenario I and according to $\lambda_s(t) = 0.01 \exp(b_s)$, $b_s \sim N(0, 0.25)$, in scenario II. In both scenarios, we took $\beta = -0.3$, $\boldsymbol{\gamma} = (0, -0.15)^\top$, and $\boldsymbol{\alpha} = (-0.3, -0.05, 0.2)^\top$. For each simulated dataset, we fit a correctly-specified stratified proportional hazards using the methods in Section 3.2 with initial values $\boldsymbol{\beta} = \mathbf{0}$ and $\lambda_{sr} = 1/\rho_s$ $(r = 1, \ldots, \rho_s; s = 1, \ldots, S)$ and with $\epsilon_{\text{tol}} = 0.0001$. To conduct the numerical approximations required for variance estimation, we set $h_M = cn^{-1/2}$ for $n = \sum_{i=1}^{M} n_i$ and $c \in \{0.1, 1, 10\}$.

Table 1 summarizes the results under scenario I with $M = 100$ and scenario II with $M = 15$; results under $M = 200$, $M = 20$, and $800 \leqslant n_i \leqslant 1000$ are similar and are displayed in Table S.1 in Web Appendix C.2. Our maximum composite likelihood procedure accurately estimated β, $\boldsymbol{\gamma}$, and $\boldsymbol{\alpha}$ in all simulation settings considered: the parameter estimators had small to negligible bias, particularly in models (10) and (12), and the extent of this bias generally diminished as the total sample size increased, whether as a result of increasing M or increasing n^*. Figure 1 displays the estimated baseline survival functions, $\widehat{S}_s(t) := \exp\left(-\sum_{\tau_{sr} \leqslant t} \widehat{\lambda}_{sr}\right)$, under scenario I with $M = 100$; these estimated curves reasonably captured the true stratum-specific baseline survival functions, $S_s(t) = \exp\{-\Lambda_s(t)\}$, regardless of the complexity of $\Lambda_1(t), \ldots, \Lambda_S(t)$ or of the parametric component of the model.

Turning to the question of inference for models (10)–(12), we see that our profile composite likelihood variance estimator produced unbiased standard error estimates for $\widehat{\beta}$, $\widehat{\boldsymbol{\gamma}}$, and $\widehat{\boldsymbol{\alpha}}$ when $M \in \{100, 200\}$; the corresponding 95% Wald-type confidence intervals achieved nominal empirical coverage under both the exchangeable and hierarchical within-cluster correlation structures (Table 1; Table S.1). Under scenario II, where $M \in \{15, 20\}$, the variance estimator remained close to unbiased, though on balance it slightly underestimated the true variability in $\widehat{\beta}$, $\widehat{\boldsymbol{\gamma}}$, and $\widehat{\boldsymbol{\alpha}}$. The resulting 95% Wald-type confidence intervals also slightly under-covered, though they still achieved above 90% empirical coverage for all regression parameters across all simulation settings. Furthermore, the magnitude of this under-coverage was similar under both the exchangeable and hierarchical correlation structures. This is notable because the composite likelihood function in (4) is the correctly-specified full-data likelihood in the within-cluster independence setting and the profile variance estimator has previously been established as valid in this context (Murphy and van der Vaart, 2000). This suggests that the slight under-coverage we observe is the result of the small sample setting—and of the distributional results in Section 3.3 not necessarily holding for $M \in \{15, 20\}$—rather than a feature of the profile composite likelihood variance estimator itself.

[Table 1 about here.]

[Figure 1 about here.]

Finally, we note that—while the variance estimation results for models (10) and (12) were quite stable across the three perturbation constants considered for numerical differentiation of the profile composite likelihood function—the results for model (11) were sensitive to large values of c (Table S.2 in Web Appendix C.3). In particular, $\widehat{Var}(\widehat{\boldsymbol{\gamma}})$ exhibited instability when $h_M = 10n^{-1/2}$, and it markedly overestimated the true variability in $\widehat{\boldsymbol{\gamma}}$ under that choice of constant. This behavior appears to be a consequence of the shape of the profile composite log-likelihood function for model (11): the slope and curvature of the function in the γ_2 direction are much greater than the slope and curvature in the γ_1 direction, in which the profile composite log-likelihood function is nearly flat, and the function itself is not symmetric about either the $\gamma_1 = \widehat{\gamma}_1$ or $\gamma_2 = \widehat{\gamma}_2$ planes (Figure S.1 in Web Appendix C.3). Thus directional derivatives of the profile composite log-likelihood function for (11) are quite sensitive to the choice of direction; furthermore, errors of order $h_M = cn^{-1/2}$ that may be acceptable in the finite differences approximation to derivatives with respect to γ_2 may be too large as to be acceptable for derivatives with respect to γ_1. While the choice of $c = 1$ appears to perform well for the dataset sizes and model formulations considered here, in practice we recommend inspecting the profile composite log-likelihood function and conducting a sensitivity analysis or small simulation study to confirm the choice of h_M.

4.2 *Performance Under Infrequent and Covariate-Dependent Monitoring*

In pragmatic CRTs and other clinical trial contexts, a subject's monitoring times $\boldsymbol{Y}_{ij}$ may be infrequent relative to the length of their event time due to practical or financial constraints and the support of these monitoring times may not be dense in the support of T_{ij}. The distribution of $(K_{ij}, \boldsymbol{Y}_{ij})$ may also depend on the subject's randomization arm: in the BCPP, for example, the combination prevention package included an expansion of HIV testing services, so that individuals in the intervention communities received more frequent HIV testing. We thus conducted a second simulation study, motivated in part by our data application, to examine the performance of our estimation and inference procedures under an infrequent and covariate-dependent monitoring schedule.

We once again considered two scenarios for the clustering and stratification factors, focusing on the $(M = 100, S = 4)$ setting for scenario I and the $(M = 15, S = 15)$ setting for scenario II, and took the within-cluster correlation structure to be hierarchical throughout. To create a covariate-dependent monitoring scheme, we assumed that all subjects underwent annual monitoring at times $Y_{ijk} \sim Unif(52k-4, 52k+4)$, $k = 1, \ldots, 4$, and that, among those subjects in clusters (communities) with $X_{ij} = 1$, additional background monitoring occurred at interval $Y_{ij,k+1} - Y_{ijk} \sim Unif(12, 24)$, $k = 5, \ldots, 19$. We simulated times to event according to the three models planned for the re-analysis of the BCPP,

$$\lambda_{ij}(t|X_{ij}, Z_{ij}) = \lambda_{Z_{ij}}(t) \exp\left(\beta X_{ij}\right) \tag{13}$$

$$\lambda_{ij}\{t|X_{ij}, P_{ij}(t), Z_{ij}\} = \lambda_{Z_{ij}}(t) \exp\left\{\alpha_{11} X_{ij} + \alpha_{12} P_{ij}(t)\right\} \tag{14}$$

$$\lambda_{ij}\{t|X_{ij}, P_{ij}(t), Z_{ij}\} = \lambda_{Z_{ij}}(t) \exp\left\{\alpha_{21} X_{ij} + \alpha_{22} P_{ij}(t) + \alpha_{23} X_{ij} P_{ij}(t)\right\}, \tag{15}$$

with X_{ij} and $P_{ij}(t)$ defined as in Section 4.1 and with the stratum-specific baseline hazard functions again generated according to Harden and Kropko (2019). We took $\beta = -1.0$, $\boldsymbol{\alpha}_1 = (-1.0, -0.5)^\top$, and $\boldsymbol{\alpha}_2 = (-1.0, -0.5, 2.0)^\top$ in the scenario I simulations and $\beta = -0.30$, $\boldsymbol{\alpha}_1 = (-0.3, -0.5)^\top$, and $\boldsymbol{\alpha}_2 = (-0.30, -0.05, 0.20)^\top$ in the scenario II simulations. For each simulated dataset, we fit models (13)–(15) using the point and interval estimators from Sections 3.2 and 3.3 with $c = 1$ and initial values and tolerances set as in Section 4.1. As a point of comparison, we also fit (13)–(15) under midpoint imputation of the interval-censored failure times, which allowed us to make use of existing methods for fitting stratified proportional hazards models with time-dependent covariates to clustered right-censored data (Therneau and Grambsch, 2000).

Results are summarized in Table 2. Comparing these to the results of the primary simulation study in Table 1, we see that our proposed estimators performed comparably with respect to both point and interval estimation for the regression parameters: any observed bias in the coefficient and variance estimators was small (representing 1.6% absolute relative bias or less for the coefficient estimators), and the corresponding 95% confidence intervals obtained > 90% empirical coverage in all settings and appropriate nominal coverage in the large M setting. Notably, the bias in the estimates of $\widehat{\beta}$, $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\alpha}}_2$ was smaller under maximum composite likelihood estimation than under midpoint imputation for each data-generating mechanism considered. This contrast was most pronounced for the coefficients of the time-dependent covariates, $P_{ij}(t)$ and $X_{ij}P_{ij}(t)$, in models (14) and (15). Midpoint imputation produced estimates for these parameters that were consistently biased towards the null. This, in turn, translated into (sometimes quite profound) under-coverage of the 95% confidence intervals: under maximum composite likelihood estimation, the empirical coverage of the 95% confidence intervals for α_{23} was 93.4% in scenario I and 91.2% in scenario II; under midpoint imputation of the failure times, the empirical coverage dropped to 54.8% and 86.2%, respectively. This disparity in performance may be due to differences in how maximum composite likelihood estimation and midpoint imputation capture person-time contributions under $P_{ij}(t) = 1$. Suppose, for example, that subject j in cluster i is interval-censored at $(L_{ij}, U_{ij}]$ and that υ_{ij} occurs in the second half of this interval. Under maximum composite likelihood estimation, this subject contributes information about risk when $P_{ij}(t) = 0$ for $t \in (L_{ij}, \upsilon_{ij})$ and risk when $P_{ij}(t) = 1$ for $t \in [\upsilon_{ij}, U_{ij}]$. If, however, this subject's failure time is midpoint imputed at $T_{ij}^{MI} := (L_{ij} + U_{ij})/2 < \upsilon_{ij}$, the information regarding $P_{ij}(t) = 1$ will be lost.

[Table 2 about here.]

Figures S.2 and S.3 in Web Appendix C.4 illustrate estimation of the stratum-specific baseline survival functions under $(M = 100, S = 4)$ and $(M = 15, S = 15)$, respectively. As expected, the estimated functions reasonably capture the true stratum-specific survival on those intervals of time to which the monitoring process assigns a non-zero probability of inspection, but does not consistently estimate the truth on those intervals with zero probability, e.g., for $t \in [0, 12)$.

5. Quantifying the Impact of Universal Test-and-Treat Adoption in Botswana

We return now to the BCPP, a CRT of combination HIV prevention strategies that was introduced and described in detail in Section 1. The primary analysis of the BCPP found evidence that combination prevention was associated with a reduction in HIV incidence in Botswana ($\widehat{HR} = 0.69$; $P = 0.09$ by permutation test; 95% CI, 0.46 to 0.90 by Cox model), but did not adjust for Botswana's mid-trial adoption of a UTT policy—a policy decision that materially changed the national standard of care for HIV testing and treatment and that may have affected the study's final significance (Makhema *et al.*, 2019).

At the time that Botswana's UTT initiative was announced, the BCPP had been underway for 2.5 years, with all 30 communities enrolled and randomized and with two of these communities (one matched pair) having completed all study follow-up (Figure 2). For those individuals still under follow-up on the announcement date, this program launch occurred at internal times ranging from 192 days to 779 days due to staggered study entry at the cluster and individual level. Exploratory analysis suggests that the trajectory of the trial may have changed following UTT adoption. Under midpoint imputation of the event times, 48 individuals in the standard-of-care arm and 31 individuals in the combination prevention arm contracted HIV by the announcement date (corresponding to approximate incidence rates of 0.0095 and 0.0062 cases/person-year, respectively); over the subsequent two years, the standard-of-care arm reported 42 new HIV diagnoses (approximate incidence rate, 0.0089 cases/person-year) while the combination prevention arm reported 26 (approximate incidence rate, 0.0056 cases/person-year).

[Figure 2 about here.]

We apply the methods of Sections 3.2 and 3.3 to a formal re-analysis of the BCPP, focusing on the impact of Botswana's changing national treatment guidelines on both the final study conclusions and on the population-level incidence of HIV more broadly. To that end, we consider three models for the time to HIV seroconversion: model (13), which estimates the main effect of combination prevention (X_{ij}) only; model (14), which estimates the main effects of both combination prevention and UTT adoption ($P_{ij}(t)$); and model (15), which allows for their possible interaction. For each of these models, we take

$$\lambda_{\boldsymbol{Z}_{ij}}(t) = \lambda_0(t) \exp\left(\boldsymbol{\eta}^\top \boldsymbol{Z}_{ij}\right), \tag{16}$$

where $\boldsymbol{Z}_{ij}$ indicates matched pair membership; we also consider the marginal counterparts of (13)–(15), in which $\boldsymbol{\eta} \overset{set}{=} \boldsymbol{0}$. We assume (i) that matched pairs of communities are independent of one another and (ii) that individual communities are independent of one another conditional on matched pair membership. We thus treat each matched pair as a distinct cluster when obtaining variance estimates for the unadjusted models and each individual community as a distinct cluster when obtaining variance estimates for the pair-stratified models. We set $h_M = n^{-1/2}$ throughout, though we also consider a sensitivity analysis with $h_M = 5n^{-1/2}$ and $h_M = 10n^{-1/2}$ for the marginal models.

Remark 4: While the fixed effect terms in (16) still permit the baseline hazard functions to differ from one pair of BCPP communities to another, they make stronger parametric assumptions about the manner in which these functions vary than would a fully non-parametric stratified baseline hazard specification (as was considered in Section 4). This choice is necessitated by identifiability concerns. For the coefficients $\boldsymbol{\beta}$ of a stratified proportional hazards model to be identifiable, the corresponding covariates must vary *within* levels of the stratification factors. As shown in Figure 2, one matched pair completed all follow-up prior to the adoption of UTT, so that $P_{ij}(t) = 0 \; \forall t \in [0, U^*_{ij}]$ for all observations in this matched pair. We thus cannot stratify on matched pair membership and still identify the policy effect of Botswana's new HIV treatment guidelines.

Results are presented in Table 3 and the estimated cumulative incidence functions in each matched pair under model (15) are given in Figure 3. Towards addressing our first research question, we find that randomization to combination HIV prevention was associated with a significant reduction in the rate of new HIV cases in Botswana (adjusted $\widehat{HR} = 0.639$; 95% CI, 0.473 to 0.862), an effect that replicated the original findings of the BCPP and that persisted even after adjusting for Botswana's new national treatment guidelines and the resulting changes to the study interventions (models (14) and (15), respectively).

The implications for our second research question are less clear. UTT implementation was a post-randomization event, such that any estimated policy effect may be subject to confounding—most notably by matched pair membership itself. The trial communities were matched into pairs on the basis of demographic factors (such as population age structure and pre-existing health infrastructure) that were thought to affect the observed within-stratum incidence rate (Makhema *et al.*, 2019; see also Figure 3); community enrollment in the BCPP was also highly staggered, such that there was large variability as to whether (and for how long) each pair of communities contributed follow-up time under the new HIV treatment guidelines (Figure 2). This intuition regarding confounding by matched pair is borne out in Table 3. Under model (15) without matched pair adjustment, UTT adoption was associated with a pronounced increase in HIV incidence in the standard of care communities, though there was high uncertainty about this estimated association (unadjusted $\widehat{HR} = 1.745$; 95% CI, 0.645 to 4.718). After accounting for matched pair membership, however, we instead find that UTT was associated with a small estimated reduction in HIV incidence (adjusted $\widehat{HR} = 0.879$; 95% CI, 0.404 to 1.913), though there remained great uncertainty about this point estimate and the estimated reduction itself was not statistically significant ($P = 0.745$).

Finally, we note that—while the estimated variance of the policy effect did vary slightly with the choice of perturbation constant for numerical differentiation—the final conclusions in Table 3 were robust to the choice of constant (Table S.3 in Web Appendix D).

[Table 3 about here.]

[Figure 3 about here.]

6. Discussion

In this manuscript, we focused on the marginal analysis of clustered and interval-censored data under the semiparametric proportional hazards framework, with immediate extensions to both stratified proportional hazards models and proportional hazards models with time-dependent covariates. To permit estimation of these models under arbitrarily complex within-cluster correlation structures and nonparametric estimation of the baseline hazard function(s), we developed a composite EM algorithm that (i) did not require specifying any second-order or higher-order moments of the data and that (ii) resulted in computationally-efficient closed-form updating equations for the baseline hazard estimator. We then established the asymptotic properties of the resulting maximum composite likelihood estimators: drawing on results from empirical process theory and semiparametric M-estimation theory, we demonstrated the consistency of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$ and the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ as the number of clusters $M \to \infty$. To facilitate inference for $\boldsymbol{\beta}$, we also developed a robust, sandwich-type variance estimator based on the Godambe information of the profile composite likelihood function. We motivated this estimator by deriving the asymptotic quadratic expansion of the profile composite log-likelihood function and extending the Taylor expansion results of Murphy and van der Vaart (2000) to the independence composite likelihood setting.

We found that our composite EM algorithm and robust variance estimator performed well across a range of simulated data settings. They yielded point estimates with near-negligible bias under all data-generating mechanisms considered—even when M was small or when the monitoring schedule was infrequent and covariate-dependent—and yielded interval estimates with the appropriate nominal coverage when M was sufficiently large and the asymptotic normal approximation for $\widehat{\boldsymbol{\beta}}$ reasonably held. We did observe some sensitivity to the choice of perturbation constant for numerical differentiation in the robust variance estimator, particularly in the model incorporating time-varying covariate effects. While $h_M = n^{-1/2}$ generally performed well in our simulation studies and data application, we nevertheless recommend visually inspecting the profile composite log-likelihood function and conducting a sensitivity analysis or small simulation study to validate the choice of h_M.

As previously noted, both the composite EM algorithm and the profile composite likelihood variance estimator require specification of the individual marginal time-to-event distributions *only*, and so do not impose any assumptions or restrictions on the within-cluster correlation structure. This robustness makes our work uniquely well-suited to the analysis of the BCPP and other infectious disease treatment and prevention CRTs, wherein the dependence between subjects may be informed by transmission along existing sexual, social, or injection drug use networks and is thus difficult to model directly. However, our model formulation and estimation approach are sufficiently broad as to be useful in a number of other important contexts. For example, stepped-wedge CRTs (SW-CRTs) are a modification of traditional cluster-randomized designs in which all clusters begin the trial in the standard-of-care arm and sequentially cross over to the intervention until all clusters are treated (Hemming *et al.*, 2015). Treatment assignment is thus a time-dependent covariate, and model (1) would be a natural choice for the monitoring and analysis of closed-cohort SW-CRTs and other crossover trials with interval-censored time-to-event endpoints. Our proposed estimation and inference procedures also have important applications to research conducted during the coronavirus 2019 (COVID-19) pandemic. The NIH Collaboratory recently published a set of guidelines for addressing COVID-19 impacts on research studies (NIH Collaboratory, 2021); these guidelines included modifying the primary analysis to adjust for the stage of the pandemic (as defined, e.g., by local COVID-19 dynamics or by impacts on research conduct) and evaluating for possible treatment effect heterogeneity (i.e., interaction effects) by pandemic stage. In open-cohort trials with a short duration of follow-up, pandemic stage might naturally be viewed as a stratification factor, while in closed-cohort trials with a longer duration of participant follow-up, pandemic stage might instead be conceived of as a time-dependent covariate. In either event, our maximum composite likelihood framework facilitates fitting these models for studies with clustered, interval-censored outcomes.

While our proposed methods reasonably estimate $\boldsymbol{\beta}$ across a range of true marginal and joint data-generating distributions, the corresponding variance estimates for $\widehat{\boldsymbol{\beta}}$ tend to be slightly anti-conservative in the small M setting. A similar phenomenon has been noted for the generalized estimating equation sandwich variance estimator and several corrections have been proposed to address this small-sample bias (e.g., Kauermann and Carroll, 2001; Mancl and DeRouen, 2001). It would be useful to derive similar small-sample corrections in the current setting.

Finally, as noted in Section 2 and explored further in the simulation studies in Section 4, a broad class of additive time-varying covariate effects may be (i) reformulated as time-dependent covariate processes and then (ii) readily estimated using our proposed composite EM algorithm. However, this reformulation requires making parametric assumptions about the form of $\boldsymbol{\beta}(t)$ that may not be tenable or verifiable; this is particularly the case for the marginal analysis of pragmatic CRTs, in which the intervention effect may vary naturally over time due to gradual intervention roll-out, or may respond in unpredictable ways to external policy decisions and forces. In light of this, extending our composite EM algorithm to consider marginal proportional hazards models with nonparametric estimation of $\boldsymbol{\beta}(t)$ (for example, through either kernel density estimation or local linear approximations) would present a promising avenue for more flexible monitoring and analysis of CRTs.

Acknowledgements

We thank the participants and study team of the Botswana Combination Prevention Project, as well as the associate editor and two referees, whose comments and suggestions substantially improved the quality of this paper. This research was supported by grants F31 AI141030, T32 AI007358, and R01 AI136947 from the U.S. National Institute of Allergy and Infectious Diseases. The BCPP Impact Evaluation was supported by the President's Emergency Plan for AIDS Relief through the Centers for Disease Control and Prevention (CDC) (cooperative agreements U01 GH000447 and U2G GH001911). The BCPP was approved by the Botswana Health Research and Development Committee and the Institutional Review Board of the Centers for Disease Control and Prevention. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official positions of the BCPP funding agencies.

Data Availability Statement

The data that support the findings of this study are available upon reasonable request from the corresponding author with approval from the BCPP study team.

References

Cai, T., Wei, L.J., and Wilcox, M. (2000). Semiparametric regression analysis for clustered failure time data. *Biometrika* **87,** 867–878.

Chandler, R.E. and Bate, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika* **94,** 167–183.

Chen, M.-H., Tong, X., and Zhu, L. (2013). A linear transformation model for multivariate interval-censored failure time data. *Canadian Journal of Statistics* **41,** 275–290.

Cook, R.J. and Tolusso, D. (2009). Second-order estimating equations for the analysis of clustered current status data. *Biostatistics* **10,** 756–772.

Gao, F., Zeng, D., Couper, D., and Lin, D.Y. (2019). Semiparametric regression analysis of multiple right- and interval-censored events. *Journal of the American Statistical Association* **114,** 1232–1240.

Gao, X. and Song, P. X.-K. (2011). Composite likelihood EM algorithm with applications to multivariate hidden Markov model. *Statistica Sinica* **21,** 165–185.

Goggins, W.B. and Finkelstein, D.M. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics* **56,** 940–943.

Goggins, W.B., Finkelstein, D.M., Schoenfeld, D.A., and Zaslavsky, A.M. A Markov chain

Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics* **54,** 1498–1507.

Harden, J.J. and Kropko, J. (2019). Simulating duration data for the Cox model. *Political Science Research and Methods* **7,** 921–928.

Hayes, R.J. and Moulton L.H. (2017). *Cluster Randomised Trials.* Boca Raton: Chapman & Hall/CRC.

Hemming, K., Haines, T.P., Chilton, P.J., Girling, A.J., and Lilford, R.J. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* **350,** h391.

Kauermann, G. and Carroll, R.J. (2001). A note on the efficiency of sandwich covariance matrix estimators. *Journal of the American Statistical Association* **96,** 1387–1396.

Kim, M.Y. and Xue, X. (2002). The analysis of multivariate interval-censored survival data. *Statistics in Medicine* **21,** 3715–3726.

Kor, C.-T., Cheng, K.-F., and Chen, Y.-H. (2013). A method for analyzing clustered interval-censored data based on Cox's model. *Statistics in Medicine* **32,** 822–832.

Kosorok, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* New York: Springer.

Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine* **13,** 2233–2247.

Makhema, J., Wirth, K.E., Holme, M.P., Gaolathe, T., Mmalane, M., Kadima, E. *et al.* (2019). Universal testing, expanded treatment, and incidence of HIV infection in Botswana. *New England Journal of Medicine* **381,** 230–242.

Mancl, L.A. and DeRouen, T.A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57,** 126–134.

Murphy, S. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American*

Statistical Association **95,** 449–465.

NIH Pragmatic Trials Collaboratory (2021). *Statistical Analysis Plan Checklist for Addressing COVID-19 Impacts.* https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/COVID-19%20SAP%20Checklist.pdf

Satten, G.A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* **83,** 355–370.

Therneau, T.M. and Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model.* New York: Springer.

Tong, X., Chen, M.-H., and Sun J. (2008) Regression analysis of multivariate interval-censored failure time data with application to tumorigenicity experiments. *Biometrical Journal* **50,** 364–374.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21,** 5–42.

Wei, L.-J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84,** 1065–1073.

World Health Organization (2016). *Botswana Launches Treat All Strategy.* https://www.afro.who.int/news/botswana-launches-treat-all-strategy.

Xu, X. and Reid, N. (2011). On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference* **141,** 3047–3054.

Yang, D., Du, M., and Sun, J. (2021). Semiparametric regression analysis of clustered interval-censored failure time data with a cured subgroup. *Statistics in Medicine* **40,** 6918–6930.

Zeng, D., Gao, F., and Lin, D.Y. (2017). Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. *Biometrika* **104,** 505–525.

Zhang, X. and Sun, J. (2010). Regression analysis of clustered interval-censored failure time data with informative cluster size. *Computational Statistics & Data Analysis* **54,** 1817–1823.

Zhang, X. and Sun, J. (2013). Semiparametric regression analysis of clustered interval-censored failure time data with informative cluster size. *The International Journal of Biostatistics* **9,** 205–214.

Supporting Information

Web Appendices referenced in Sections 3.2, 3.3, 4, and 5 are available with this paper at the Biometrics website on Wiley Online Library. All code needed to implement the composite EM algorithm and profile composite likelihood variance estimator is publicly available at https://github.com/kaitlyncook/composite-ic-cox.

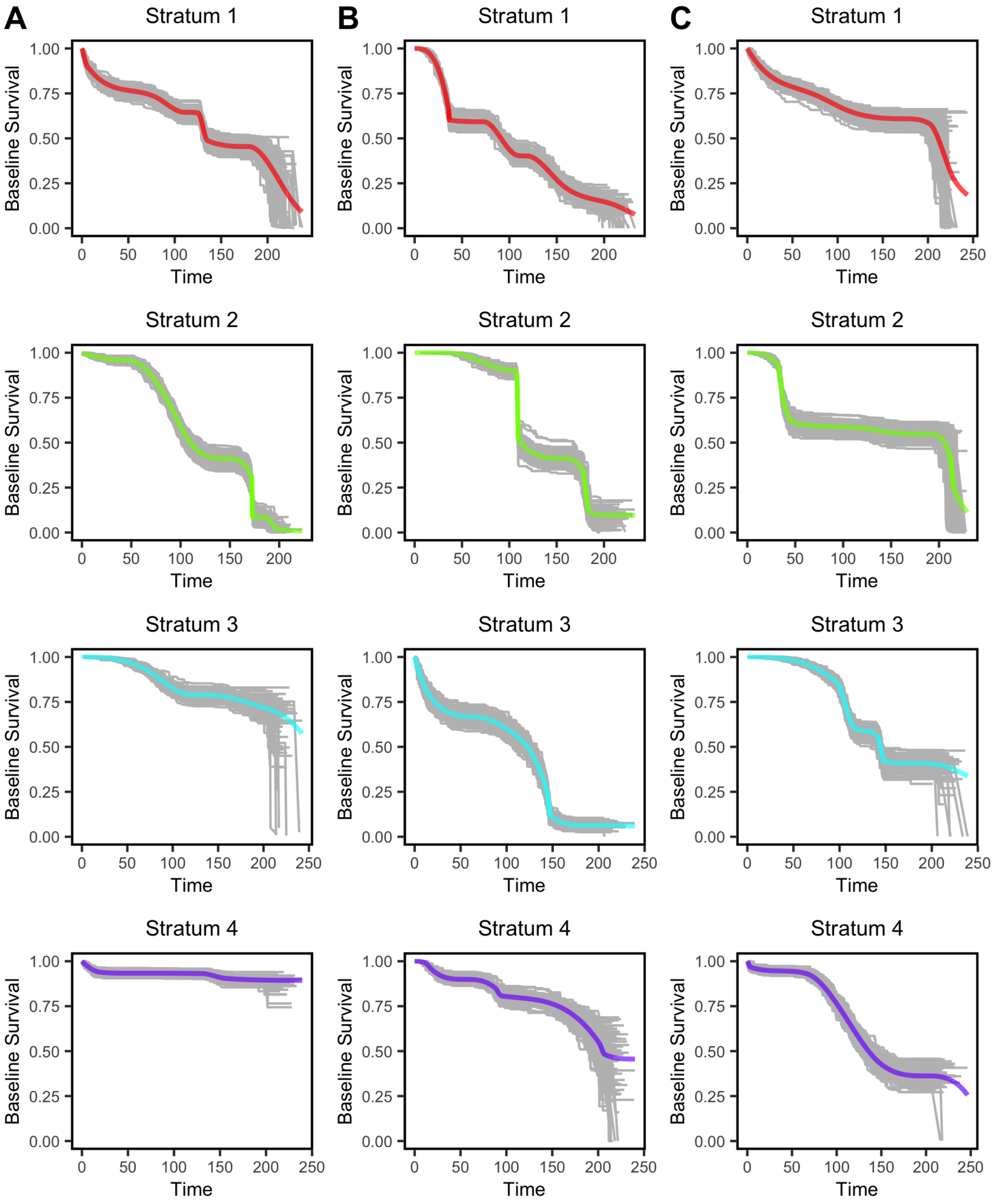


Figure 1. Comparison of 100 randomly selected estimated stratum-specific baseline survival functions (in gray) with the true data-generating functions (in color) under model (10) (column A), model (11) (column B), and model (12) (column C) with $M = 100$ and $S = 4$.

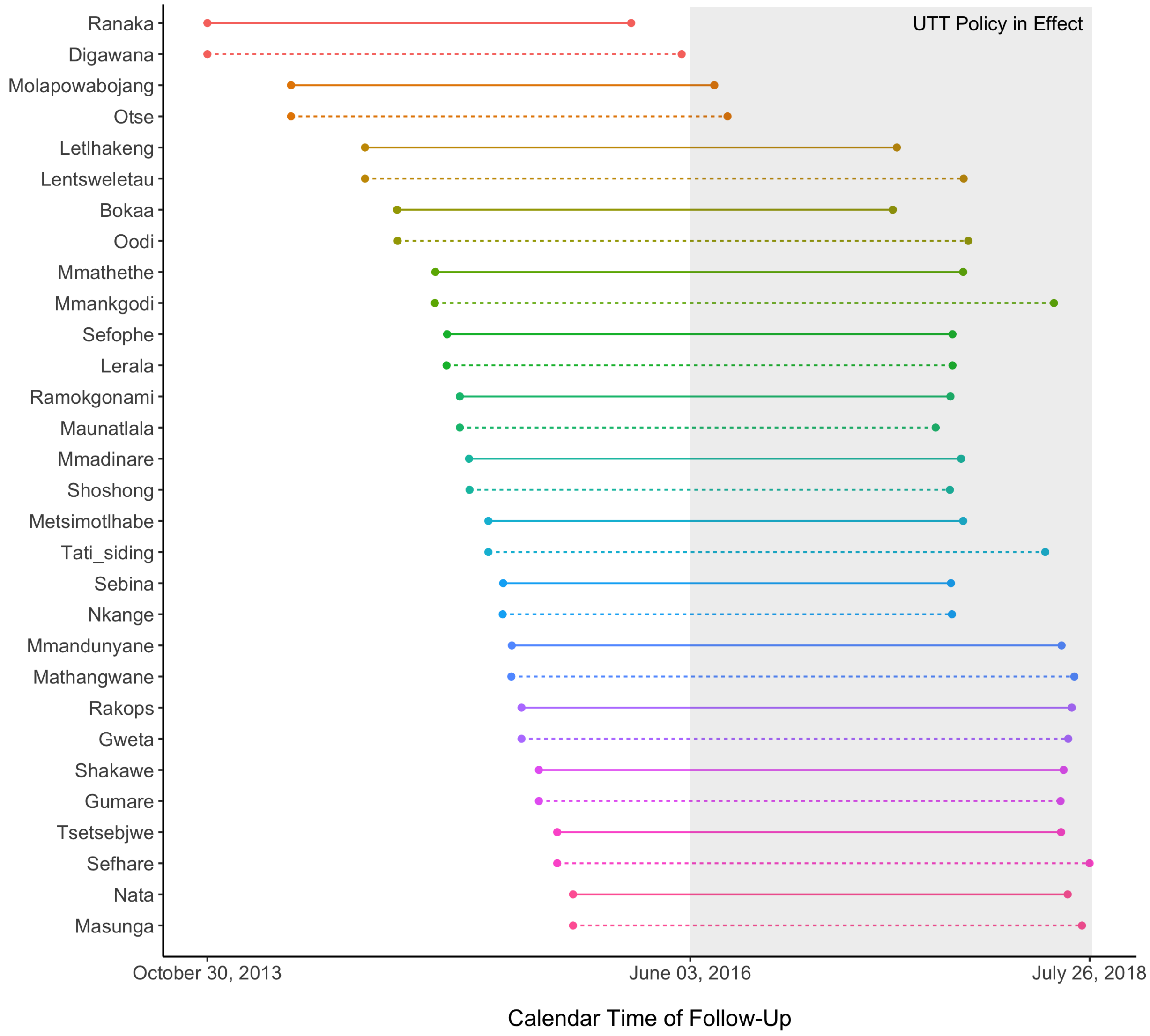


Figure 2. Duration of follow-up for each community in the Botswana Combination Prevention Project, measured as the time from the earliest recorded study visit to the last recorded study visit across all community members. Intervention assignment is indicated by line type (standard of care, solid; combination prevention, dashed) and pair membership by color; the shaded region corresponds to the time during which the national universal-test-and-treat (UTT) policy was in effect.

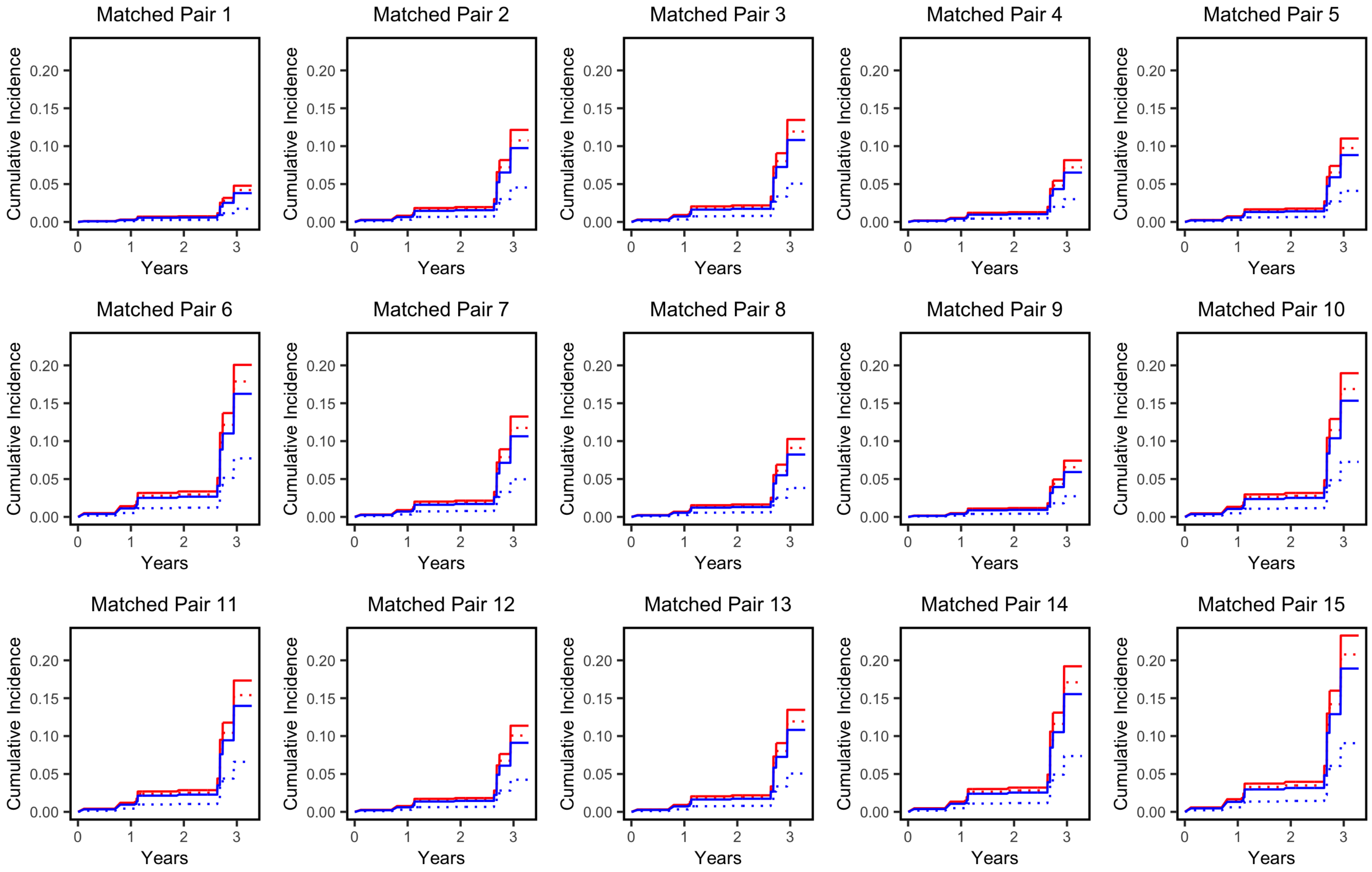


Figure 3. Estimated cumulative probability of HIV seroconversion in the standard-of-care communities (in red) and combination prevention communities (in blue), both before (solid line) and after (dotted line) universal test-and-treat adoption in Botswana.

Table 1

Finite sample performance of the maximum composite likelihood estimators and profile composite likelihood variance estimators under $h_M = n^{-1/2}$.

	Within-Cluster Independence					Copula Dependence Model				
	Point Est.	Bias	Emp. SE	Est. SE	CP	Point Est.	Bias	Emp. SE	Est. SE	CP
Scenario I: $S = 4$, $M = 100$, $20 \leqslant n_i \leqslant 30$										
Model (10)										
$\beta =$ -0.30	-0.303	-0.003	0.061	0.063	95.3%	-0.305	-0.005	0.112	0.107	93.3%
Model (11)										
$\gamma_1 =$ 0.00	0.034	0.034	0.153	0.156	94.8%	0.027	0.027	0.285	0.289	95.0%
$\gamma_2 =$ -0.15	-0.158	-0.008	0.040	0.039	94.5%	-0.156	-0.006	0.063	0.064	95.2%
Model (12)										
$\alpha_1 =$ -0.30	-0.301	-0.001	0.083	0.082	94.1%	-0.304	-0.004	0.117	0.116	93.7%
$\alpha_2 =$ -0.05	-0.049	0.001	0.136	0.138	94.6%	-0.049	0.001	0.098	0.098	95.5%
$\alpha_3 =$ 0.20	0.197	-0.003	0.175	0.179	95.1%	0.205	0.005	0.132	0.136	95.7%
Scenario II: $S = 15$, $M = 15$, $500 \leqslant n_i \leqslant 700$										
Model (10)										
$\beta =$ -0.30	-0.301	-0.001	0.026	0.024	90.9%	-0.302	-0.002	0.047	0.045	93.0%
Model (11)										
$\gamma_1 =$ 0.00	0.025	0.025	0.089	0.094	95.0%	0.031	0.031	0.124	0.123	93.3%
$\gamma_2 =$ -0.15	-0.157	-0.007	0.024	0.025	95.5%	-0.159	-0.009	0.031	0.032	93.5%
Model (12)										
$\alpha_1 =$ -0.30	-0.301	-0.001	0.030	0.027	91.8%	-0.305	-0.005	0.050	0.046	90.4%
$\alpha_2 =$ -0.05	-0.049	0.001	0.045	0.043	91.5%	-0.051	-0.001	0.047	0.044	91.9%
$\alpha_3 =$ 0.20	0.200	0.000	0.059	0.056	91.6%	0.203	0.003	0.065	0.062	91.5%

Point Est., empirical average of the parameter estimator; Bias, empirical average of the bias; Emp. SE, empirical standard error; Est. SE, empirical average of the standard error estimator; CP, empirical coverage probability of the corresponding 95% Wald-type confidence interval. All results are summarized across 1000 simulation replicates.

Table 2

Finite sample performance of the maximum composite likelihood estimators and profile composite likelihood variance estimators relative to midpoint imputation in the Botswana Combination Prevention Project setting.

	Maximum Composite Likelihood					Midpoint Imputation				
	Point Est.	Bias	Emp. SE	Est. SE	CP	Point Est.	Bias	Emp. SE	Est. SE	CP
Scenario I: $S = 4$, $M = 100$, $20 \leqslant n_i \leqslant 30$										
Model (13)										
β = -1.00	-0.996	0.004	0.103	0.099	94.4%	-0.986	0.014	0.105	0.100	94.4%
Model (14)										
α_{11} = -1.00	-1.011	-0.011	0.105	0.102	95.2%	-0.962	0.038	0.103	0.103	93.6%
α_{12} = -0.50	-0.500	0.000	0.074	0.076	96.4%	-0.438	0.062	0.066	0.070	86.6%
Model (15)										
α_{21} = -1.00	-1.016	-0.016	0.111	0.106	93.6%	-1.047	-0.047	0.111	0.105	91.4%
α_{22} = -0.50	-0.507	-0.007	0.092	0.094	94.8%	-0.480	0.020	0.093	0.095	95.0%
α_{23} = 2.00	2.028	0.028	0.128	0.125	93.4%	1.786	-0.214	0.121	0.120	54.8%
Scenario II: $S = 15$, $M = 15$, $500 \leqslant n_i \leqslant 700$										
Model (13)										
β = -0.30	-0.302	-0.002	0.049	0.047	91.2%	-0.287	0.013	0.049	0.051	93.6%
Model (14)										
α_{11} = -0.30	-0.303	-0.003	0.051	0.048	91.2%	-0.327	-0.027	0.052	0.063	94.6%
α_{12} = -0.50	-0.503	-0.003	0.035	0.034	92.6%	-0.458	0.042	0.034	0.032	72.8%
Model (15)										
α_{21} = -0.30	-0.304	-0.004	0.053	0.051	91.4%	-0.239	0.061	0.053	0.056	79.4%
α_{22} = -0.05	-0.050	0.000	0.049	0.047	91.8%	-0.028	0.022	0.044	0.043	90.6%
α_{23} = 0.20	0.201	0.001	0.068	0.062	91.2%	0.158	-0.042	0.063	0.059	86.2%

Point Est., empirical average of the parameter estimator; Bias, empirical average of the bias; Emp. SE, empirical standard error; Est. SE, empirical average of the standard error estimator; CP, empirical coverage probability of the corresponding 95% Wald-type confidence interval. All results are summarized across 500 simulation replicates.

Table 3

Analysis of the Botswana Combination Prevention Project, accounting for both randomization to combination HIV prevention and the mid-trial adoption of a national universal test-and-treat policy (UTT).

	Marginal Model			Adjusted for Pair Membership		
	$\widehat{HR}$	95% CI	P-value	$\widehat{HR}$	95% CI	P-value
Model (13)						
Combination prevention	0.629	(0.430, 0.921)	0.017	0.639	(0.473, 0.862)	0.003
Model (14)						
Combination prevention	0.629	(0.426, 0.930)	0.020	0.638	(0.473, 0.861)	0.003
Universal test-and-treat adoption	1.173	(0.490, 2.807)	0.721	0.661	(0.014, 30.350)	0.832
Model (15)						
Combination prevention						
Prior to UTT adoption	0.871	(0.467, 1.624)	0.664	0.792	(0.569, 1.102)	0.166
Post UTT adoption	0.392	(0.179, 0.861)	0.020	0.408	(0.180, 0.925)	0.032
Universal test-and-treat adoption	1.745	(0.645, 4.718)	0.273	0.879	(0.404, 1.913)	0.745

$\widehat{HR}$, estimated hazard ratio $\exp(\widehat{\beta})$; CI, confidence interval.

Supporting Information for Marginal Proportional Hazards Models for Clustered Interval-Censored Data with Time-Dependent Covariates

by Kaitlyn Cook[1,*], Wenbin Lu[2,**], and Rui Wang[1,3,**]

[1]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, Massachusetts, U.S.A.

[2]Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A.

[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, U.S.A

[*] *email:* kaitlyn_cook@harvardpilgrim.org

[**] *email:* wlu4@ncsu.edu

[***] *email:* rwang@hsph.harvard.edu

Web Appendix A. Composite Expectation Maximization under the Independence Composite Likelihood

A.1 Review of Existing Literature on Composite Expectation Maximization

Suppose that the data consist of M independent clusters (indexed by $i = 1, \dots, M$), with n_i possibly dependent observations in cluster i (indexed by $j = 1, \dots, n_i$). Let $\boldsymbol{\mathcal{O}}$ denote the full collection of observed data, and suppose that the univariate margins of its joint distribution are parametrized by $\theta \in \Theta$. Then the independence composite likelihood function for θ is given by

$$\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}}) = \prod_{i=1}^{M} \prod_{j=1}^{n_i} f(\mathcal{O}_{ij}; \theta). \tag{S.1}$$

To facilitate maximization of (S.1), suppose that the observed data collection for observation j in cluster i is (or may be conceived as) a many-to-one function of some augmented data vector, $\mathcal{A}_{ij}$, so that

$$f(\mathcal{O}_{ij}; \theta) = \int_{\{\mathcal{A}_{ij} : \mathcal{O}(\mathcal{A}_{ij}) = \mathcal{O}_{ij}\}} f(\mathcal{A}_{ij}; \theta) \mu(d\mathcal{A}_{ij}). \tag{S.2}$$

We then define the following working composite models for the observed, augmented, and conditional data distributions, with each composite model formed by multiplying together a collection of congenial univariate densities:

$$f^*(\boldsymbol{\mathcal{O}}; \theta) := \prod_{i=1}^{M} \prod_{j=1}^{n_i} f(\mathcal{O}_{ij}; \theta)$$

$$f^*(\boldsymbol{\mathcal{A}}; \theta) := \prod_{i=1}^{M} \prod_{j=1}^{n_i} f(\mathcal{A}_{ij}; \theta)$$

$$f^*(\boldsymbol{\mathcal{A}} | \boldsymbol{\mathcal{O}}; \theta) := \prod_{i=1}^{M} \prod_{j=1}^{n_i} f(\mathcal{A}_{ij} | \mathcal{O}_{ij}; \theta),$$

where $f(\mathcal{A}_{ij} | \mathcal{O}_{ij}; \theta) = f(\mathcal{A}_{ij}; \theta) / f(\mathcal{O}_{ij}; \theta)$. Then the composite EM algorithm identifies stationary points of (S.1) by iteratively (i) constructing the objective function

$$\mathcal{Q}_C(\theta | \theta_l) = \mathbb{E}^*_{\theta_l}[\log f^*(\boldsymbol{\mathcal{A}}; \theta) | \boldsymbol{\mathcal{O}}] = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathbb{E}_{\theta_l}[\log f(\mathcal{A}_{ij}; \theta) | \mathcal{O}_{ij}],$$

where $\mathbb{E}^*_{\theta_l}[\cdot|\boldsymbol{\mathcal{O}}]$ and $\mathbb{E}_{\theta_l}[\cdot|\mathcal{O}_{ij}]$ denote expectations taken with respect to $f^*(\boldsymbol{\mathcal{A}}|\boldsymbol{\mathcal{O}};\theta_l)$ and $f(\mathcal{A}_{ij}|\mathcal{O}_{ij};\theta_l)$, respectively, and then (ii) updating $\theta_{l+1} = \text{argmax}\mathcal{Q}_C(\theta|\theta_l)$.

This independence composite EM algorithm retains all of the key attributes of the traditional EM procedure, namely the ascent property (reprinted as Theorem S.1 below for reference) and the convergence of the algorithm to a stationary point of the independence composite likelihood (reprinted as Theorem S.2 below for reference). These results are special cases of the results presented in Gao and Song (2011) for more general composite EM algorithms, and the structure of their proofs is identical to those in Dempster *et al.* (1977) and Wu (1983) for traditional EM algorithms, though with the joint distributions of $\boldsymbol{\mathcal{O}}$, $\boldsymbol{\mathcal{A}}$, and $\boldsymbol{\mathcal{A}}|\boldsymbol{\mathcal{O}}$ now replaced by their working composite models.

Theorem S.1: *Let θ_l and θ_{l+1} be successive iterations of the composite EM algorithm. Then $\log\mathcal{L}_C(\theta_{l+1};\boldsymbol{\mathcal{O}}) \geqslant \log\mathcal{L}_C(\theta_l;\boldsymbol{\mathcal{O}})$ for all $l \in \mathbb{N}$, with equality if and only if both $\mathcal{Q}_C(\theta_{l+1}|\theta_l) = \mathcal{Q}_C(\theta_l|\theta_l)$ and $f^*(\boldsymbol{\mathcal{A}}|\boldsymbol{\mathcal{O}};\theta_{l+1}) = f^*(\boldsymbol{\mathcal{A}}|\boldsymbol{\mathcal{O}};\theta_l)$ a.e.*

Theorem S.2: *Let $(\theta_l)_{l\in\mathbb{N}}$ be a sequence of composite EM updates, and suppose that*

i. $\Theta_0 = \{\theta : \log\mathcal{L}_C(\theta;\boldsymbol{\mathcal{O}}) \geqslant \log\mathcal{L}_C(\theta_0;\boldsymbol{\mathcal{O}})\}$ *is compact for all* $\log\mathcal{L}_C(\theta_0;\boldsymbol{\mathcal{O}}) > -\infty$

ii. $\log\mathcal{L}_C(\theta;\boldsymbol{\mathcal{O}})$ *is continuous in θ and differentiable on the interior of* Θ

iii. $\mathcal{Q}_C(\theta|\theta_l)$ *is continuous in both θ and θ_l.*

Then the sequence $(\log\mathcal{L}_C(\theta_l;\boldsymbol{\mathcal{O}}))_{l\in\mathbb{N}}$ converges monotonically to $\ell^* = \log\mathcal{L}_C(\theta^*;\boldsymbol{\mathcal{O}})$ for some stationary point, θ^*. Let $\mathscr{S}(\ell^*)$ be the set of all stationary points for which $\log\mathcal{L}_C(\theta;\boldsymbol{\mathcal{O}}) = \ell^*$. If $\mathscr{S}(\ell^*)$ contains only a single point, then it also follows that $\theta_l \to \theta^*$.

Note that the ascent property guarantees the independence composite likelihood for the observed data is monotone non-decreasing across each iteration of the composite EM algorithm, while Theorem S.2 establishes sufficient conditions under which the sequence of independence composite EM updates converges to some stationary point θ^* of (S.1). As with the classical EM algorithm, there is no general guarantee that the limiting θ^* for any given realization of the composite EM algorithm is equal to the global maximizer, $\widehat{\theta}_{CL}$, unless the composite likelihood function is strictly concave.

A.2 Poisson Data Augmentation Produces a Composite EM Algorithm

Although the composite likelihood function, $\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$, and the corresponding composite EM objective function, $\mathcal{Q}_C(\theta|\theta_l)$, are derived as either the product or the sum of independent cluster-level contributions, the adoption of a working independence model for these cluster-level contributions implies that all terms in $\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$ and $\mathcal{Q}_C(\theta|\theta_l)$ are exchangeable and thus may be rearranged and reindexed. In the main manuscript, we rearrange $\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$ and $\mathcal{Q}_C(\theta|\theta_l)$ as a matter of convenience so that observations are partitioned into strata (determined by levels of the stratification factors, $\boldsymbol{Z}_{ij}$) as opposed to clusters (determined by i):

$$\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}}) = \prod_{i=1}^{M}\prod_{j=1}^{n_i} f(\mathcal{O}_{ij}; \theta) \equiv \prod_{s=1}^{S}\prod_{v=1}^{n_s} f(\mathcal{O}_{sv}; \theta)$$

$$\mathcal{Q}_C(\theta|\theta_l) = \sum_{i=1}^{M}\sum_{j=1}^{n_i} \mathbb{E}_{\theta_l}[\log f(\mathcal{A}_{ij}; \theta)|\mathcal{O}_{ij}] \equiv \sum_{s=1}^{S}\sum_{v=1}^{n_s} \mathbb{E}_{\theta_l}[\log f(\mathcal{A}_{sv}; \theta)|\mathcal{O}_{sv}],$$

where $s = 1, \ldots, S$ indexes stratum membership and $v = 1, \ldots, n_s$ indexes study subjects within stratum s. In what follows, we will establish that Poisson augmentation leads to a composite EM algorithm using cluster-level indexing of $\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$ and $\mathcal{Q}_C(\theta|\theta_l)$. Given that these arguments rely only on (i) the continuity of $\mathcal{Q}_C(\theta|\theta_l)$, (ii) the boundedness of $\log \mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$, and (iii) the nature of the coarsening relationship between $\mathcal{A}_{ij}$ and $\mathcal{O}_{ij}$—all of which are trivially preserved under reindexing from (i, j) to (s, v)—they will also hold for the iterative estimation procedure proposed in Section 3.2 of the main manuscript.

We now verify that Poisson augmentation does indeed produce an independence composite EM algorithm satisfying the conditions of Theorems S.1 and S.2 above. To that end, we note that the observed data for individual j in cluster i is given by

$$\mathcal{O}_{ij} = \{(L_{ij}, U_{ij}], \boldsymbol{Z}_{ij}, (\boldsymbol{X}_{ij}(t) : t \leqslant U^*_{ij})\},$$

where $(L_{ij}, U_{ij}]$ is the censoring interval containing the true time to event T_{ij}, $\boldsymbol{X}_{ij}(t)$ is a p-dimensional bounded covariate process informing the time-to-event distribution, and $\boldsymbol{Z}_{ij}$ is a set of stratification factors taking on $S \in \mathbb{N}$ distinct levels (denoted by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_S$). Then the individual observed data density function under the marginal stratified proportional hazards model in equation (1) of the main text,

$$\lambda_{ij}\{t|\boldsymbol{X}_{ij}(t), \boldsymbol{Z}_{ij}\} = \lambda_{\boldsymbol{Z}_{ij}}(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)\}, \tag{S.3}$$

is given by

$$\begin{aligned} f(\mathcal{O}_{ij}; \theta) &= \exp\left\{-\int_0^{L_{ij}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_{\boldsymbol{Z}_{ij}}(t)\right\} - \exp\left\{-\int_0^{U_{ij}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_{\boldsymbol{Z}_{ij}}(t)\right\} \\ &= \prod_{s=1}^{S} \left[\exp\left\{-\int_0^{L_{ij}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t)\right\} - \exp\left\{-\int_0^{U_{ij}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t)\right\}\right]^{I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s)}. \end{aligned}$$

To simplify notation, we assume without loss of generality that $\boldsymbol{Z}_{ij} = \boldsymbol{z}_s$. Then under nonparametric estimation of the baseline hazard function Λ_s, which restricts Λ_s to the class of step functions $\Lambda_s(t) = \sum_{\tau_{sr} \leqslant t} \lambda_{sr}$ with non-negative jumps λ_{sr} at times τ_{sr} $(r = 1, \ldots, \rho_s)$, we may write

$$f(\mathcal{O}_{ij}; \theta) = \exp\left(-\sum_{\tau_{sr} \leqslant L_{ij}} \lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right) \left\{1 - \exp\left(-\sum_{L_{ij} < \tau_{sr} \leqslant U_{ij}} \lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)\right\}^{I(U_{ij} < \infty)}$$

for $\boldsymbol{X}_{ijr} \equiv \boldsymbol{X}_{ij}(\tau_{sr})$.

To maximize this expression, we introduce for each subject a collection of independent Poisson random variables, $W_{ijr} \sim Pois(\lambda_{sr} \exp\{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}\})$ $(r = 1, \ldots, \rho_s)$. The augmented data collection for individual j in cluster i is then

$$\mathcal{A}_{ij} = \{(L_{ij}, U_{ij}], \boldsymbol{Z}_{ij}, (\boldsymbol{X}_{ij}(t) : t \leqslant U^*_{ij}), \{W_{ijr} : r = 1 \ldots, \rho_s; \tau_{sr} \leqslant U^*_{ij}\}\}$$

and the associated augmented data density is given by

$$f(\mathcal{A}_{ij};\theta)=\prod_{r=1}^{\rho_s}\left\{\frac{1}{W_{ijr}!}\left(\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)^{W_{ijr}}\exp\left(-\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)\right\}^{I(\tau_{sr}\leqslant U_{ij}^*)}.$$

We first establish that the coarsening relationship in (S.2) holds for $\mathcal{A}_{ij}$ and $\mathcal{O}_{ij}$ in the stratified proportional hazards setting. To that end, we define $A_{ij} := \sum_{\tau_{sr}\leqslant L_{ij}} W_{ijr}$ and $B_{ij} := I(U_{ij} < \infty)\sum_{L_{ij}<\tau_{sr}\leqslant U_{ij}} W_{ijr}$, and note that $(L_{ij}, U_{ij}]$ is represented by the event $\{A_{ij}=0\}\cap\{B_{ij}=0\}$ when $U_{ij}=\infty$ and by the event $\{A_{ij}=0\}\cap\{B_{ij}>0\}$ when $U_{ij}>\infty$. Considering first the case when $U_{ij}=\infty$, we have that

$$\begin{aligned}
f(\mathcal{O}_{ij};\theta) &\overset{?}{=} \int_{\{\mathcal{A}_{ij}:\mathcal{O}(\mathcal{A}_{ij})=\mathcal{O}_{ij}\}} f(\mathcal{A}_{ij};\theta)\mu(d\mathcal{A}_{ij})\\
&= \sum_{\{\boldsymbol{W}_{ij}:A_{ij}=0\}}\prod_{r=1}^{\rho_s}\left\{\frac{1}{W_{ijr}!}\left(\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)^{W_{ijr}}\exp\left(-\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)\right\}^{I(\tau_{sr}\leqslant U_{ij}^*)}\\
&= \sum I(A_{ij}=0)\times\prod_{r=1}^{\rho_s}\left\{\frac{1}{W_{ijr}!}\left(\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)^{W_{ijr}}\exp\left(-\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)\right\}^{I(\tau_{sr}\leqslant U_{ij}^*)}\\
&= P(A_{ij}=0)\\
&\overset{\checkmark}{=} \exp\Big(-\sum_{\tau_{sr}\leqslant L_{sk}}\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\Big),
\end{aligned}$$

where the last equality follows from the observation that $A_{ij}\sim Pois\{\sum_{\tau_{sr}\leqslant L_{ij}}\lambda_{sr}\exp(\boldsymbol{\beta}^\top\boldsymbol{X}_{ijr})\}$. Similarly, when $U_{ij}<\infty$, we find

$$\begin{aligned}
f(\mathcal{O}_{ij};\theta) &\overset{?}{=} \int_{\{\mathcal{A}_{ij}:\mathcal{O}(\mathcal{A}_{ij})=\mathcal{O}_{ij}\}} f(\mathcal{A}_{ij};\theta)\mu(d\mathcal{A}_{ij})\\
&= \sum_{\{\boldsymbol{W}_{ij}:A_{ij}=0\cap B_{ij}>0\}}\prod_{r=1}^{\rho_s}\left\{\frac{1}{W_{ijr}!}\left(\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)^{W_{ijr}}\exp\left(-\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)\right\}^{I(\tau_{sr}\leqslant U_{ij}^*)}\\
&= \sum I(A_{ij}=0\cap B_{ij}>0)\times\prod_{r=1}^{\rho_s}\left\{\frac{1}{W_{ijr}!}\left(\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)^{W_{ijr}}\exp\left(-\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right)\right\}^{I(\tau_{sr}\leqslant U_{ij}^*)}\\
&= P(A_{ij}=0)P(B_{ij}>0)\\
&\overset{\checkmark}{=} \exp\Big(-\sum_{\tau_{sr}\leqslant L_{ij}}\lambda_{sr}e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\Big)\Big\{1-\exp\Big(-\sum_{L_{ij}<\tau_{sr}\leqslant U_{ij}}\lambda_{sr}e^{\boldsymbol{\beta}\top \boldsymbol{X}_{ijr}}\Big)\Big\},
\end{aligned}$$

where the second-to-last equality follows from the independence of $\{W_{ijr} : \tau_{sr} \leqslant L_{ij}\}$ and

$\{W_{ijr} : L_{ij} < \tau_{sr} \leqslant U_{ij}\}$, and the last equality follows from the observation that $B_{ij} = I(U_{ij} < \infty) \sum_{L_{ij}<\tau_{sr}\leqslant U_{ij}} W_{ijr} \equiv \sum_{L_{ij}<\tau_{sr}\leqslant U_{ij}} W_{ijr} \sim Pois\{\sum_{L_{ij}<\tau_{sr}\leqslant U_{ij}} \lambda_{sr} \exp(\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr})\}$. So the coarsening relationship in (S.2) holds with respect to the proposed augmentation variables, and, given the congeniality of the marginal distributions for $\mathcal{O}_{ij}$, $\mathcal{A}_{ij}$, and $\mathcal{A}_{ij}|\mathcal{O}_{ij}$, the estimation procedure in Section 3.2 of the main text is a composite EM algorithm satisfying Theorem S.1.

We next consider Theorem S.2, which concerns the convergence of the composite EM algorithm to some stationary point θ^* of the independence composite log-likelihood function. The composite EM objective function is

$$\mathcal{Q}_C(\theta|\theta_l) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left[\sum_{r=1}^{\rho_s} I(\tau_{sr} \leqslant U_{ij}^*) \left\{\widehat{W}_{ijr} \log\left(\lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right) - \lambda_{sr} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ijr}}\right\}\right],$$

with

$$\widehat{W}_{ijr} = \mathbb{E}_{\theta_l}[W_{ijr}|\mathcal{O}_{ij}] = \begin{cases} \frac{\lambda_{l,sr}\exp\left(\boldsymbol{\beta}_l^\top \boldsymbol{X}_{ijr}\right)}{1-\exp\left\{-\sum_{L_{ij}<\tau_{sr}\leqslant U_{ij}} \lambda_{l,sr}\exp\left(\boldsymbol{\beta}_l^\top \boldsymbol{X}_{ijr}\right)\right\}} & \text{for } L_{ij} < \tau_{sr} \leqslant U_{ij}^* \\ 0 & \text{otherwise} \end{cases}.$$

Given that $\mathcal{Q}_C(\theta|\theta_l)$ is continuous in both θ and θ_l; that $\log \mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}})$ is a bounded, continuous, and differentiable function of θ; and under the assumption that the stratified proportional hazards model is identifiable under the independence composite likelihood construction (see Condition 6 in Web Appendix B.1), the proposed Poisson augmentation algorithm also satisfies the conditions of Theorem S.2 and so its iterates $\theta_l \to \theta^*$. While this limiting θ^* is guaranteed to be a stationary point of the composite log-likelihood function, we do not have a general guarantee that it will be the global maximum.

A.3 Projection onto the Working Composite Observed Data Model

To evaluate the projection $\mathbb{E}_{\theta_l}[W_{ijr}|\mathcal{O}_{ij}]$ for those $r = 1, \ldots, \rho_s$ such that $\tau_{sr} \leqslant U_{ij}^*$, we first note that $U_{ij}^* = L_{ij} I(U_{ij} = \infty) + U_{ij} I(U_{ij} < \infty)$ represents the total time during which subject j in cluster i is under monitoring for the event of interest. For all observations, this active monitoring time includes the interval $(0, L_{ij}]$; for those observations with $U_{ij} < \infty$, it also includes the interval $(L_{ij}, U_{ij}]$. We thus consider two separate cases for $\mathbb{E}_{\theta_l}[W_{ijr}|\mathcal{O}_{ij}]$: (i) when the corresponding $\tau_{sr} \leqslant L_{ij}$, and (ii) when the corresponding $L_{ij} < \tau_{sr} \leqslant U^*_{ij}$.

(i) A necessary condition for L_{ij} to be the left endpoint of the observed censoring interval is that no event occurs prior to that time, i.e., that $A_{ij} = 0$, which occurs if and only if $W_{ijr} = 0$ for all $\tau_{sr} \leqslant L_{ij}$. Thus $\mathbb{E}_{\theta_l}[W_{ijr}|\mathcal{O}_{ij}] = 0$ for all $\tau_{sr} \leqslant L_{ij}$.

(ii) Consider now $L_{ij} < \tau_{sr} \leqslant U^*_{ij}$, which is a non-empty subset of $\boldsymbol{\tau}_s$ only when $U^*_{ij} = U_{ij} < \infty$. For these observations, we note that the observed censoring interval $(L_{ij}, U_{ij}]$ is equivalent to the event $\{A_{ij} = 0\} \cap \{B_{ij} > 0\}$, so that

$$\begin{aligned}\mathbb{E}_{\theta_l}[W_{ijr}|\boldsymbol{X}_{ij}(t)] &= \mathbb{E}_{\theta_l}[W_{ijr}|A_{ij} = 0, \boldsymbol{X}_{ij}(t)] \\ &= \mathbb{E}_{\theta_l}[W_{ijr}|A_{ij} = 0, B_{ij} > 0, \boldsymbol{X}_{ij}(t)]P(B_{ij} > 0|\boldsymbol{X}_{ij}(t)) \\ &\quad + \mathbb{E}_{\theta_l}[W_{ijr}|A_{ij} = 0, B_{ij} = 0, \boldsymbol{X}_{ij}(t)]P(B_{ij} = 0|\boldsymbol{X}_{ij}(t)) \\ &= \mathbb{E}_{\theta_l}[W_{ijr}|\mathcal{O}_{ij}]P(B_{ij} > 0|\boldsymbol{X}_{ij}(t)) + 0.\end{aligned}$$

Rearranging, we find

$$\mathbb{E}_{\theta_l}[W_{ijr}|\mathcal{O}_{ij}] = \frac{\mathbb{E}_{\theta_l}[W_{ijr}|\boldsymbol{X}_{ij}(t)]}{P(B_{ij} > 0|\boldsymbol{X}_{ij}(t))} = \frac{\lambda_{l,sr} \exp(\boldsymbol{\beta}_l^\top \boldsymbol{X}_{ijr})}{1 - \exp\{\sum_{L_{ij} < \tau_{sr} \leqslant U_{ij}} \lambda_{l,sr} \exp(\boldsymbol{\beta}_l^\top \boldsymbol{X}_{ijr})\}}.$$

The expression shown in Section 3.2 of the main text follows as a result of reindexing the data with respect to stratum and subjects within stratum, (s, v), as opposed to cluster and subjects within cluster, (i, j).

Web Appendix B. Asymptotic Results

B.1 Setting and Regularity Conditions

Let $Y_{ij1} < \cdots < Y_{ij,K_{ij}}$ be the sequence of monitoring times for subject j in cluster i, and set $Y_{ij0} = 0$, $Y_{ij,K_{ij}+1} = \infty$, and $\delta_{ijk} = I(Y_{ijk} < T_{ij} \leqslant Y_{ij,k+1})$ for $k = 0, \ldots, K_{ij}$. Then the observed censoring interval $(L_{ij}, U_{ij}]$ is that interval among $\{(Y_{ijk}, Y_{ij,k+1}] : k = 0, \ldots, K_{ij}\}$

for which $\delta_{ijk} = 1$, and the full-data independence composite likelihood function for (S.3) may be written as

$$\begin{aligned}
\mathcal{L}_C(\theta; \boldsymbol{\mathcal{O}}) &= \prod_{i=1}^{M} \prod_{j=1}^{n_i} f(\mathcal{O}_{ij}; \theta) \\
&= \prod_{i=1}^{M} \prod_{j=1}^{n_i} \prod_{s=1}^{S} \left[\exp\left\{ -\int_0^{L_{ij}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} - \exp\left\{ -\int_0^{U_{ij}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \right]^{I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s)} \\
&= \prod_{i=1}^{M} \prod_{j=1}^{n_i} \prod_{s=1}^{S} \left(\prod_{k=0}^{K_{ij}} \left[\exp\left\{ -\int_0^{Y_{ijk}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \right.\right. \\
&\qquad\qquad \left.\left. - \exp\left\{ -\int_0^{Y_{ij,k+1}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \right]^{\delta_{ijk}} \right)^{I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s)} .
\end{aligned}$$

The independence composite log-likelihood function is then given by

$$\begin{aligned}
\ell_C(\theta; \boldsymbol{\mathcal{O}}) = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) &\left(\sum_{k=0}^{K_{ij}} \delta_{ijk} \log \left[\exp\left\{ -\int_0^{Y_{ijk}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \right.\right. \\
&\qquad \left.\left. - \exp\left\{ -\int_0^{Y_{ij,k+1}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \right] \right),
\end{aligned}$$

and the maximum composite likelihood estimator under nonparametric estimation of the baseline cumulative hazard functions, $\widehat{\theta}_{CL} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$ with $\widehat{\boldsymbol{\Lambda}} = (\widehat{\Lambda}_1, \ldots, \widehat{\Lambda}_S)$, satisfies

$$\widehat{\theta}_{CL} = \underset{\theta \in \mathcal{B} \times \mathcal{C}}{\operatorname{argmax}}\ \ell_C(\theta; \boldsymbol{\mathcal{O}}), \tag{S.4}$$

where $\mathcal{B} \subset \mathbb{R}^p$ and $\mathcal{C} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_S$ with $\mathcal{C}_s$ the set of step functions with non-negative jumps at times τ_{sr} $(r = 1, \ldots, \rho_s)$.

We establish the asymptotic properties of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$ under the setting in which the number of clusters $M \to \infty$ and assuming the following regularity conditions, adapted from Zeng *et al.* (2017) and Gao *et al.* (2019):

Condition 1: The maximum cluster size $n = n(M) = \max_{1 \leqslant i \leqslant M} n_i$ is bounded above by a positive constant n^* for all M and n_i is independent of $\{T_{ij} : j = 1, \ldots, n_i\}$ and $\{K_{ij}, \boldsymbol{Y}_{ij} : j = 1, \ldots, n_i\}$ conditional on $\{\boldsymbol{Z}_{ij}, \boldsymbol{X}_{ij}(t) : j = 1, \ldots, n_i\}$.

Condition 2: The number of potential monitoring times K_{ij} is positive with $\mathbb{E}(K_{ij}) < \infty$, and the monitoring times themselves have finite support $\mathcal{Y}$ with least upper bound τ. In addition, there exists some positive constant ξ such that $P\{\min_{0\leqslant k<K_{ij}}(Y_{ij,k+1} - Y_{ijk}) \geqslant \xi | K_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}\} = 1$. Finally, there exists a probability measure μ on $\mathcal{Y}$ such that the bivariate distribution function of $(Y_{ijk}, Y_{ij,k+1})$ conditional on $(K_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij})$ is dominated by $\mu \times \mu$ and its Radon-Nikodym derivative, denoted by $\tilde{f}_k(u, v; K_{ij}, \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij})$, can be expanded to a positive and twice-continuously differentiable function in the set $\{(u, v) : 0 \leqslant u \leqslant \tau, 0 \leqslant v \leqslant \tau, v - u \geqslant \xi\}$.

Condition 3: The set of stratification factors $\boldsymbol{Z}_{ij}$ is a discrete random vector taking on finitely many distinct values, $\{\boldsymbol{z}_s : s = 1, \ldots, S\}$, with S fixed.

Condition 4: With probability 1, $\boldsymbol{X}_{ij}(\cdot)$ has bounded total variation in $\mathcal{Y}$. If there exists a constant vector $\boldsymbol{a}_1 \in \mathbb{R}^p$ and a deterministic function $a_2(t)$ such that $\boldsymbol{a}_1^\top \boldsymbol{X}_{ij}(t) + a_2(t) = 0$ for any $t \in \mathcal{Y}$ with probability 1, then $\boldsymbol{a}_1 = \boldsymbol{0}$ and $a_2(t) = 0$ for any $t \in \mathcal{Y}$.

Condition 5: The true value of $\boldsymbol{\beta}$, denoted by $\boldsymbol{\beta}_0$, lies in the interior of a known compact set $\mathcal{B} \subset \mathbb{R}^p$. For $s = 1, \ldots, S$, the true value $\Lambda_{s0}(\cdot)$ of $\Lambda_s(\cdot)$ is strictly increasing and continuously differentiable on $[0, \tau]$ with $\Lambda_{s0}(0) = 0$.

Condition 6: If there exists $\boldsymbol{\beta}_* \in \mathcal{B}$ and strictly increasing and continuously differentiable $\Lambda_{s*}(t)$ for $s = 1, \ldots, S$ and $t \in \mathcal{Y}$ with $\Lambda_{s*}(0) = 0$ such that

$$\prod_{j=1}^{n_i}\prod_{s=1}^{S}\left(\sum_{k=0}^{K_{ij}} \delta_{ijk}\left[\exp\left\{-\int_0^{Y_{ijk}} e^{\boldsymbol{\beta}_*^\top \boldsymbol{X}_{ij}(u)} d\Lambda_{s*}(u)\right\} - \exp\left\{-\int_0^{Y_{ij,k+1}} e^{\boldsymbol{\beta}_*^\top \boldsymbol{X}_{ij}(u)} d\Lambda_{s*}(u)\right\}\right]\right)^{I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)}$$

$$= \prod_{j=1}^{n_i}\prod_{s=1}^{S}\left(\sum_{k=0}^{K_{ij}} \delta_{ijk}\left[\exp\left\{-\int_0^{Y_{ijk}} e^{\boldsymbol{\beta}_0^\top \boldsymbol{X}_{ij}(u)} d\Lambda_{s0}(u)\right\} - \exp\left\{-\int_0^{Y_{ij,k+1}} e^{\boldsymbol{\beta}_0^\top \boldsymbol{X}_{ij}(u)} d\Lambda_{s0}(u)\right\}\right]\right)^{I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)}$$

with probability 1, then $\boldsymbol{\beta}_* = \boldsymbol{\beta}_0$ and $\Lambda_{s*}(t) = \Lambda_{s0}(t)$ for $s = 1, \ldots, S$ and $t \in \mathcal{Y}$.

Let $\mathbb{P}_M$ denote the empirical measure for M independent clusters, P denote the true probability measure, and $\mathbb{G}_M = \sqrt{M}(\mathbb{P}_M - P)$ denote the corresponding empirical process. We reformulate maximum composite likelihood estimation under the independence composite likelihood as an M-estimation task in Web Appendix B.2, and then establish the consistency of $\widehat{\theta}_{CL}$ in Web Appendix B.3, the asymptotic normality of the parametric component $\widehat{\boldsymbol{\beta}}$ in Web Appendix B.4, and the form of the quadratic expansion of the profile composite log-likelihood function for $\boldsymbol{\beta}$ in Web Appendix B.5. These proofs make use of three additional lemmas, which are presented and proved in Web Appendix B.6.

B.2 Reformulation of the Maximum Composite Likelihood Estimator as a Semiparametric M-Estimator

Denote the individual cluster-level contribution to the independence composite log-likelihood function by $m(\boldsymbol{\beta}, \boldsymbol{\Lambda}; \boldsymbol{\mathcal{O}}_i) = m_\theta(\boldsymbol{\mathcal{O}}_i)$, where

$$\begin{aligned} m_\theta(\boldsymbol{\mathcal{O}}_i) := \sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \Bigg(\sum_{k=0}^{K_{ij}} \delta_{ijk} \log \Bigg[& \exp\left\{ - \int_0^{Y_{ijk}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \\ & - \exp\left\{ - \int_0^{Y_{ij,k+1}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \Bigg] \Bigg), \end{aligned} \tag{S.5}$$

and define the empirical and population criterion functions as

$$\mathbb{A}_M(\theta) := \mathbb{P}_M m_\theta \qquad \text{and} \qquad A(\theta) := P m_\theta,$$

respectively, for all $\theta \in \Theta$, where the parameter space $\Theta = \mathcal{B} \times \mathcal{D}$ with $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_S$ and $\mathcal{D}_s = \{\Lambda_s : \Lambda_s \text{ is a non-decreasing function with } \Lambda_s(0) = 0 \text{ and } \Lambda_s(\tau) < \infty\}$. We equip $\mathcal{B}$ with the standard Euclidean norm,

$$d(\boldsymbol{\beta}, \boldsymbol{\beta}_0) := \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = \sqrt{(\beta_1 - \beta_{10})^2 + \cdots + (\beta_p - \beta_{p0})^2} \qquad \text{for } \boldsymbol{\beta}, \boldsymbol{\beta}_0 \in \mathcal{B},$$

and $\mathcal{D}_s$ with the supremum norm on $\mathcal{Y}$,

$$d(\Lambda_s, \Lambda_{s0}) := \|\Lambda_s - \Lambda_{s0}\|_{l^\infty(\mathcal{Y})} = \sup\{|\Lambda_s(t) - \Lambda_{s0}(t)| : t \in \mathcal{Y}\} \qquad \text{for } \Lambda_s, \Lambda_{s0} \in \mathcal{D}_s.$$

Note that $\mathbb{A}_M(\theta)$ is proportional to $\ell_C(\theta; \boldsymbol{\mathcal{O}})$, so that the maximum composite likelihood estimator defined in (S.4) likewise maximizes $\mathbb{A}_M(\theta)$:

$$\mathbb{A}_M(\widehat{\theta}_{CL}) = \sup_{\theta \in \mathcal{B} \times \mathcal{C}} \mathbb{A}_M(\theta).$$

Thus the maximum composite likelihood estimator $\widehat{\theta}_{CL}$ is a type of semiparametric M estimator, and the asymptotic behavior of $\widehat{\theta}_{CL}$ may be understood through the lens of semiparametric M-estimation theory, with the consistency of $\widehat{\theta}_{CL}$ and the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ closely related to properties of the function class $\mathcal{M} := \{m_\theta(\cdot) : \theta \in \Theta\}$.

B.3 Consistency of the Maximum Composite Likelihood Estimator

Theorem 1: *Under Conditions 1–6,* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \sum_{s=1}^{S} \|\widehat{\Lambda}_s - \widehat{\Lambda}_{s0}\|_{l^\infty(\mathcal{Y})} \overset{a.s.}{\to} 0.$

Proof. Using Lemma S.1 and similar arguments to those in Zeng *et al.* (2017), we find that the nonparametric maximum composite likelihood estimator $\widehat{\boldsymbol{\Lambda}}$ satisfies $\limsup_M \widehat{\Lambda}_s(\tau - \epsilon) < \infty$ with probability 1 for any $\epsilon > 0$ and for any $s = 1, \ldots, S$. By choosing a decreasing sequence $\{\epsilon_n : n \in \mathbb{N}\}$ such that $\epsilon_n \downarrow 0$, we may then use Helly's selection theorem to conclude that $\widehat{\Lambda}_s$ converges pointwise to some $\Lambda_{s*} \in \mathcal{D}_s$ on any compact interior subset of $\mathcal{Y}$. Thus for M large enough we may restrict our attention to estimators $\widehat{\Lambda}_s \in \mathcal{D}_s$, $s = 1, \ldots, S$.

We next establish that the population criterion function $A(\theta)$ has a well-separated and unique maximum at θ_0. To that end, note that $m_\theta(\boldsymbol{\mathcal{O}}_i)$ may be rewritten as

$$\begin{aligned} m_\theta(\boldsymbol{\mathcal{O}}_i) &= \sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left(\sum_{k=0}^{K_{ij}} \delta_{ijk} \log \left[P\big\{ Y_{ijk} < T_{ij} \leqslant Y_{ij,k+1} | K_{ij}, \boldsymbol{Y}_{ij}, \boldsymbol{X}_{ij}(t), \boldsymbol{Z}_{ij}; \theta \big\} \right] \right) \\ &= \sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left[\sum_{k=0}^{K_{ij}} \delta_{ijk} \log \{ p_{ijk}(\theta) \} \right], \end{aligned}$$

so that the population criterion function is given by

$$\begin{aligned} A(\theta) &= P m_\theta \\ &= \mathbb{E} \left(\mathbb{E}_{\theta_0} \left[\sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \sum_{k=0}^{K_{ij}} \delta_{ijk} \log \{ p_{ijk}(\theta) \} \middle| \boldsymbol{K}_i, \boldsymbol{Y}_i, \boldsymbol{X}_i(t), \boldsymbol{Z}_i \right] \right) \\ &= \mathbb{E} \left[\sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \sum_{k=0}^{K_{ij}} \mathbb{E}_{\theta_0} \left\{ I(Y_{ijk} < T_{ij} \leqslant Y_{ij,k+1}) \big| \boldsymbol{K}_i, \boldsymbol{Y}_i, \boldsymbol{X}_i(t), \boldsymbol{Z}_i \right\} \log \{ p_{ijk}(\theta) \} \right] \end{aligned}$$

$$= \mathbb{E}\left[\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s)\sum_{k=0}^{K_{ij}} p_{ijk}(\theta_0)\log\{p_{ijk}(\theta)\}\right].$$

Without loss of generality, we assume (for the moment) that $S = 1$, so that

$$A(\theta) = \mathbb{E}\left\{\sum_{j=1}^{n_i}\sum_{k=0}^{K_{ij}} p_{0,ijk}\log\left(p_{ijk}\right)\right\}, \tag{S.6}$$

where we have used the shorthand $p_{0,ijk} = p_{ijk}(\theta_0)$ and $p_{ijk} = p_{ijk}(\theta)$. The integrand in (S.6) is the negative cross-entropy between the collection of true marginal probabilities, $\boldsymbol{p}_0 = \{p_{0,ijk} : k = 0, \ldots, K_{ij}; j = 1, \ldots, n_i\}$, and the collection of predicted marginal probabilities, $\boldsymbol{p} = \{p_{ijk} : k = 0, \ldots, K_{ij}; j = 1, \ldots, n_i\}$. This expression will then be maximized at $\boldsymbol{p} = \boldsymbol{p}_0 \iff p_{ijk} = p_{0,ijk}\ \forall j, k \implies \theta = \theta_0$, where the last implication follows from the identifiability of the stratified proportional hazards model under the independence composite likelihood construction (Condition 6). Thus θ_0 represents the unique maximizer of $A(\theta)$. Given that $A(\theta)$ is a continuous function of θ, it also follows that this maximizer is well-separated.

Finally, we note that the function class $\mathcal{M} = \{m_\theta(\cdot) : \theta \in \Theta\}$ is P-Glivenko-Cantelli given Lemma S.1. Then

$$\sup_{\theta\in\Theta}|\mathbb{P}_M m_\theta - P m_\theta| = \sup_{\theta\in\Theta}|\mathbb{A}_M(\theta) - A(\theta)| \overset{a.s.}{\to} 0,$$

and it follows from an application of the Argmax Theorem (see Kosorok (2008), Theorem 2.12) that $d(\widehat{\theta}_{CL}, \theta_0) := \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \sum_{s=1}^{S}\|\widehat{\Lambda}_s - \widehat{\Lambda}_{s0}\|_{l^\infty(\mathcal{Y})} \overset{a.s.}{\to} 0$.

B.4 Asymptotic Normality of the Parametric Component

To study the limiting distribution of the parametric component $\widehat{\boldsymbol{\beta}}$ of the maximum composite likelihood estimator $\widehat{\theta}_{CL}$, we first characterize the first and second partial derivatives of the criterion function $m_\theta(\boldsymbol{\mathcal{O}}_i)$ under the parametric submodel $\boldsymbol{\Lambda}_{\epsilon,\boldsymbol{h}}$ that satisfies the relationship

$$d\boldsymbol{\Lambda}_{\epsilon,\boldsymbol{h}} = \left((1+\epsilon h_1)d\Lambda_1, \ldots, (1+\epsilon h_S)d\Lambda_S\right)^\top$$

for $\epsilon \in \mathbb{R}$, $\boldsymbol{h} = (h_1, \ldots, h_S)$, and $h_s \in L_2(\mu)$; to simplify notation, we also introduce the following expressions:

$$\begin{aligned}
S(v; \boldsymbol{\beta}, \Lambda) &= \exp\left\{ -\int_0^v e^{\boldsymbol{\beta}^\top \boldsymbol{X}(u)} d\Lambda(u) \right\} \\
B(t, v; \boldsymbol{\beta}, \Lambda) &= -I(v \geqslant t) S(v; \boldsymbol{\beta}, \Lambda) e^{\boldsymbol{\beta}^\top \boldsymbol{X}(t)} \\
C(t, v, w; \boldsymbol{\beta}, \Lambda) &= \frac{B(t, v; \boldsymbol{\beta}, \Lambda) - B(t, w; \boldsymbol{\beta}, \Lambda)}{S(v; \boldsymbol{\beta}, \Lambda) - S(w; \boldsymbol{\beta}, \Lambda)} \\
D(t, u, v, w; \boldsymbol{\beta}, \Lambda) &= \frac{[-I(v \geqslant t) B(u, v; \boldsymbol{\beta}, \Lambda) + I(w \geqslant t) B(u, w; \boldsymbol{\beta}, \Lambda)] e^{\boldsymbol{\beta}^\top \boldsymbol{X}(t)}}{S(v; \boldsymbol{\beta}, \Lambda) - S(w; \boldsymbol{\beta}, \Lambda)}.
\end{aligned}$$

Then the composite score function for $\boldsymbol{\beta}$ is given by

$$\begin{aligned}
m_1(\boldsymbol{\beta}, \boldsymbol{\Lambda}) &:= \frac{\partial}{\partial \boldsymbol{\beta}^\top} m(\boldsymbol{\beta}, \boldsymbol{\Lambda}; \boldsymbol{\mathcal{O}}_i) \\
&= \sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left\{ \sum_{k=0}^{K_{ij}} \delta_{ijk} \int_0^\tau C(t, Y_{ijk}, Y_{ij,k+1}; \boldsymbol{\beta}, \Lambda_s) \boldsymbol{X}_{ij}(t) d\Lambda_s(t) \right\} \qquad \text{(S.7)}
\end{aligned}$$

and the composite score operator for $\boldsymbol{\Lambda}$ along the submodel $d\boldsymbol{\Lambda}_{\epsilon,\boldsymbol{h}}$ is given by

$$\begin{aligned}
m_2(\boldsymbol{\beta}, \boldsymbol{\Lambda})[\boldsymbol{h}] &:= \frac{\partial}{\partial \epsilon} m(\boldsymbol{\beta}, \boldsymbol{\Lambda}_{,\boldsymbol{h}}; \boldsymbol{\mathcal{O}}_i) \Big|_{\epsilon=0} \\
&= \sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left\{ \sum_{k=0}^{K_{ij}} \delta_{ijk} \int_0^\tau C(t, Y_{ijk}, Y_{ij,k+1}; \boldsymbol{\beta}, \Lambda_s) h_s(t) d\Lambda_s(t) \right\}. \qquad \text{(S.8)}
\end{aligned}$$

Note that, in M estimation contexts such as the one we consider here, a natural extension of (S.8) takes derivatives along the p-dimensional submodel $\boldsymbol{\Lambda}_{\boldsymbol{\epsilon},\boldsymbol{h}}$ instead, where $d\boldsymbol{\Lambda}_{\boldsymbol{\epsilon},\boldsymbol{h}} = \big((1 + \boldsymbol{\epsilon}^\top \boldsymbol{h}_1) d\Lambda_1, \ldots, (1 + \boldsymbol{\epsilon}^\top \boldsymbol{h}_S) d\Lambda_S\big)^\top$ for $\boldsymbol{\epsilon} \in \mathbb{R}^p$, $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_S)$, and $\boldsymbol{h}_s$ a p-dimensional vector of functions in $L_2(\mu)$; in what follows, whether we refer to the score operator for $\boldsymbol{\Lambda}$ under the one- or p-dimensional submodel should hopefully be clear from context.

We may similarly define

$$\begin{aligned}
m_{11}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) &:= \frac{\partial}{\partial \boldsymbol{\beta}} m_1(\boldsymbol{\beta}, \boldsymbol{\Lambda}) \qquad \text{(S.9)} \\
&= \sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left\{ \sum_{k=0}^{K_{ij}} \delta_{ijk} \int_0^\tau C(t, Y_{ijk}, Y_{ij,k+1}; \boldsymbol{\beta}, \Lambda_s) \boldsymbol{X}_{ij}(t)^{\otimes 2} d\Lambda_s(t) \right\}
\end{aligned}$$

$$
\begin{aligned}
&-\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left[\sum_{k=0}^{K_{ij}}\delta_{ijk}\left\{\int_0^\tau C(t,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)\boldsymbol{X}_{ij}(t)d\Lambda_s(t)\right\}^{\otimes 2}\right]\\
&+\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left\{\sum_{k=0}^{K_{ij}}\delta_{ij}\int_0^\tau\int_0^\tau D(t,u,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)\boldsymbol{X}_{ij}(t)\boldsymbol{X}_{ij}(u)^\top d\Lambda_s(u)d\Lambda_s(t)\right\}
\end{aligned}
$$

$$
\begin{aligned}
m_{21}(\boldsymbol{\beta},\boldsymbol{\Lambda})[\boldsymbol{h}] &:= \frac{\partial}{\partial\boldsymbol{\beta}} m_2(\boldsymbol{\beta},\boldsymbol{\Lambda})[\boldsymbol{h}] \qquad \text{(S.10)}\\
&= \sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left\{\sum_{k=0}^{K_{ij}}\delta_{ijk}\int_0^\tau C(t,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s h_s(t)\boldsymbol{X}_{ij}(t)^\top d\Lambda_s(t)\right\}\\
&\quad -\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left\{\sum_{k=0}^{K_{ij}}\delta_{ijk}\int_0^\tau\int_0^\tau C(t,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)\right.\\
&\qquad\qquad \left.\times C(u,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)h_s(t)\boldsymbol{X}_{ij}(u)^\top d\Lambda_s(u)d\Lambda_s(t)\right\}\\
&\quad +\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left\{\sum_{k=0}^{K_{ij}}\delta_{ijk}\int_0^\tau\int_0^\tau D(t,u,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)h_s(t)\boldsymbol{X}_{ij}(u)^\top d\Lambda_s(u)d\Lambda_s(t)\right\}
\end{aligned}
$$

$$
\begin{aligned}
m_{12}(\boldsymbol{\beta},\boldsymbol{\Lambda})[\boldsymbol{h}] &:= \frac{\partial}{\partial\epsilon} m_1(\boldsymbol{\beta},\boldsymbol{\Lambda}_{\epsilon,\boldsymbol{h}})\Big|_{\epsilon=0} \qquad \text{(S.11)}\\
&= \sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left\{\sum_{k=0}^{K_{ij}}\delta_{ijk}\int_0^\tau C(t,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)\boldsymbol{X}_{ij}(t)h_s(t)^\top d\Lambda_s(t)\right\}\\
&\quad -\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left\{\sum_{k=0}^{K_{ij}}\delta_{ijk}\int_0^\tau\int_0^\tau C(t,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)\right.\\
&\qquad\qquad \left.\times C(u,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)\boldsymbol{X}_{ij}(t)h_s(u)^\top d\Lambda_s(u)d\Lambda_s(t)\right\}\\
&\quad +\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left\{\sum_{k=0}^{K_{ij}}\delta_{ijk}\int_0^\tau\int_0^\tau D(t,u,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)\boldsymbol{X}_{ij}(t)h_s(u)^\top d\Lambda_s(u)d\Lambda_s(t)\right\}
\end{aligned}
$$

$$
\begin{aligned}
m_{22}(\boldsymbol{\beta},\boldsymbol{\Lambda})[\boldsymbol{h}][\widetilde{\boldsymbol{h}}] &= \frac{\partial}{\partial\epsilon} m_2(\boldsymbol{\beta},\boldsymbol{\Lambda}_{\epsilon,\tilde{\boldsymbol{h}}})[\boldsymbol{h}]\Big|_{\epsilon=0} \qquad \text{(S.12)}\\
&= \sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left\{\sum_{k=0}^{K_{ij}}\delta_{ijk}\int_0^\tau C(t,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)h_s(t)\widetilde{h}_s(t)^\top d\Lambda_s(t)\right\}\\
&\quad -\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left\{\sum_{k=0}^{K_{ij}}\delta_{ijk}\int_0^\tau\int_0^\tau C(t,Y_{ijk},Y_{ij,k+1};\boldsymbol{\beta},\Lambda_s)\right.
\end{aligned}
$$

$$\left.\times C(u, Y_{ijk}, Y_{ij,k+1}; \boldsymbol{\beta}, \Lambda_s) h_s(t) \widetilde{h}_s(u)^\top d\Lambda_s(u) d\Lambda_s(t) \right\}$$

$$+ \sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left\{ \sum_{k=0}^{K_{ij}} \delta_{ijk} \int_0^\tau \int_0^\tau D(t, u, Y_{ijk}, Y_{ij,k+1}; \boldsymbol{\beta}, \Lambda_s) h_s(t) \widetilde{h}_s(u)^\top d\Lambda_s(u) d\Lambda_s(t) \right\}.$$

Note that (S.11) and (S.12) may be extended in a similar fashion to (S.8) to consider derivatives along the p-dimensional submodel with $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_S)$ and $\widetilde{\boldsymbol{h}} = (\widetilde{\boldsymbol{h}}_1, \ldots, \widetilde{\boldsymbol{h}}_S)$ for $\boldsymbol{h}_s$, $\widetilde{\boldsymbol{h}}_s$ p-dimensional vectors of functions in $L_2(\mu)$.

Remark 1: From our construction of $\widehat{\theta}_{CL}$ as the maximizer of $\mathbb{A}_M(\theta) = \mathbb{P}_M m_\theta$, it also follows that

$$\mathbb{P}_M m_1(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) = \boldsymbol{0} \quad \text{and} \quad \mathbb{P}_M m_2(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}] = 0$$

for all $\boldsymbol{h} = (h_1, \ldots, h_S)$ with $h_s \in L_2(\mu)$.

In order to establish the weak convergence of $\sqrt{M}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, we make two additional assumptions regarding the behavior and properties of $m_\theta(\boldsymbol{\mathcal{O}}_i)$ under P:

Condition 7: There exists some $\boldsymbol{h}^* = (\boldsymbol{h}_1^*, \ldots, \boldsymbol{h}_S^*)$ with $\boldsymbol{h}_s^*$ a p-dimensional vector of functions in $L_2(\mu)$ such that, for any $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_S)$ with $\boldsymbol{h}_s$ a p-dimensional vector of functions in $L_2(\mu)$,

$$\mathbb{E}_{\theta_0}(m_{12}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}] - m_{22}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*][\boldsymbol{h}]) = \boldsymbol{0}. \tag{S.13}$$

Furthermore, each element of $\boldsymbol{h}_s^*$ $(s = 1, \ldots, S)$ can be expanded to be a continuously differentiable function in $[0, \tau]$ with bounded total variation.

Condition 8: The matrix $\mathbb{E}_{\theta_0}\{m_{11}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_{21}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\}$ is symmetric and invertible, and the matrix

$$\begin{aligned} I^* = {} & \mathbb{E}_{\theta_0}\left\{m_{11}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_{21}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}^{-1} \\ & \times \mathbb{E}_{\theta_0}\left[\{m_1(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_2(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\}^{\otimes 2}\right] \end{aligned} \tag{S.14}$$

$$\times \mathbb{E}_{\theta_0} \left\{ m_{11}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_{21}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*] \right\}^{-1},$$

satisfies $0 < \det(I^*) < \infty$.

Remark 2: Equation (S.13) in Condition 7 is analogous to the projection of the score equation for $\boldsymbol{\beta}$ onto the closed linear span of the nuisance score operators for $\boldsymbol{\Lambda}$ in the semiparametric maximum likelihood estimation context and its solution $\boldsymbol{h}^*$ is analogous to the corresponding least favorable direction (cf. Chapter 25 of Van der Vaart (2000)). This $\boldsymbol{h}^*$ will exist provided that $m_{22}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\cdot][\boldsymbol{h}]$ is a bounded linear operator with bounded inverse. That each element of $\boldsymbol{h}_s^*$, $s = 1, \ldots, S$, may be expanded into a continuously differentiable function with bounded total variation on $[0, \tau]$ is necessary for establishing Glivenko-Cantelli and Donsker results. Condition 8 guarantees the existence, finiteness, and non-singularity of the asymptotic variance for $\widehat{\boldsymbol{\beta}}$.

Theorem 2: *Under Conditions 1–8, $\sqrt{M}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges weakly to a p-dimensional zero-mean normal random vector with covariance matrix I^*.*

Proof. We begin by noting that, given Condition 7 and by similar arguments to those in Lemma S.2, the function classes

$$\mathcal{M}_1 := \{m_1(\boldsymbol{\beta}, \boldsymbol{\Lambda}) : \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\Lambda} \in \mathcal{D}^*\}$$

and

$$\mathcal{M}_2 := \{m_2(\boldsymbol{\beta}, \boldsymbol{\Lambda})[\boldsymbol{h}^*] : \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\Lambda} \in \mathcal{D}^*\}$$

are both P-Donsker. As a result, the associated empirical processes $\mathbb{G}_M \{m_1(\boldsymbol{\beta}, \boldsymbol{\Lambda})\}$ and $\mathbb{G}_M \{m_2(\boldsymbol{\beta}, \boldsymbol{\Lambda})[\boldsymbol{h}^*]\}$ are both stochastically equicontinuous, with

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \leqslant \delta_M,\ d(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}_0) \leqslant C^* M^{-1/3}} \left| \sqrt{M}(\mathbb{P}_M - P)\{m_1(\boldsymbol{\beta}, \boldsymbol{\Lambda}) - m_1(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\} \right| = o_P(1)$$

and

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\leqslant\delta_M,\ d(\boldsymbol{\Lambda},\boldsymbol{\Lambda}_0)\leqslant C^*M^{-1/3}} \left|\sqrt{M}(\mathbb{P}_M - P)\{m_2(\boldsymbol{\beta},\boldsymbol{\Lambda})[\boldsymbol{h}^*] - m_2(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\}\right| = o_P(1)$$

for any $\delta_M \downarrow 0$ and any $C^* > 0$. Furthermore, by Theorem 1, $\widehat{\theta}_C = (\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})$ is consistent for $\theta_0 = (\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)$, so that for M large enough we may find $C < \infty$ such that $\widehat{\Lambda}_s(\tau) \leqslant C$ for $s = 1,\ldots,S$. Thus $m_1(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})$ and $m_2(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}^*]$ belong to $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, and we may write

$$\begin{aligned}
\mathbb{G}_M\left\{m_1(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})\right\} &= \mathbb{G}_M\left\{m_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right\} + o_P(1)\\
\implies -\sqrt{M}P\left\{m_1(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})\right\} &= \sqrt{M}(\mathbb{P}_M - P)\left\{m_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right\} + o_P(1)\\
\implies \sqrt{M}P\left\{m_1(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) - m_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right\} &= -\sqrt{M}\mathbb{P}_M\left\{m_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right\} + o_P(1) \qquad \text{(S.15)}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{G}_M\left\{m_2(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}^*]\right\} &= \mathbb{G}_M\left\{m_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\} + o_P(1)\\
\implies -\sqrt{M}P\left\{m_2(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}^*]\right\} &= \sqrt{M}(\mathbb{P}_M - P)\left\{m_2(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\} + o_P(1)\\
\implies \sqrt{M}P\left\{m_2(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}^*] - m_2(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\} &= -\sqrt{M}\mathbb{P}_M\left\{m_2(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\} + o_P(1).
\end{aligned}$$

$$\text{(S.16)}$$

We take Taylor series expansions of the left-hand sides of (S.15) and (S.16) about $(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)$, using the results of Lemma S.3 to bound the second-order terms. In particular, we have

$$\begin{aligned}
&P\left\{m_1(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}) - m_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right\}\\
&\quad= P\left\{m_{11}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right\}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + P\left\{m_{12}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0]\right\}\\
&\qquad+ \mathbb{E}_{\theta_0}\left\{O(1)\left(\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s)\left[\sum_{k=0}^{K_{ij}}\left\{\widehat{\Lambda}_s(Y_{ijk}) - \Lambda_{s0}(Y_{ijk})\right\}^2\right]\right) + O(1)\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2\right\}\\
&\quad= \mathbb{E}_{\theta_0}\left\{m_{11}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right\}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \mathbb{E}_{\theta_0}\left\{m_{12}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0]\right\} + O_P\left(M^{-2/3} + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2\right),
\end{aligned}$$

where $m_{12}(\boldsymbol{\beta},\boldsymbol{\Lambda})[\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0]$ denotes the derivative of $m_1(\boldsymbol{\beta},\boldsymbol{\Lambda})$ along the submodel $d\boldsymbol{\Lambda}_0 + \epsilon d(\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0)$, which is given by equation (S.11) with $h_s(u)d\Lambda_s(u)$ replaced by $d(\widehat{\Lambda}_s - \Lambda_{s0})(u)$, and

$$
\begin{aligned}
&P\left\{m_2(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}})[\boldsymbol{h}^*]-m_2(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}\\
&\quad=P\left\{m_{21}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+P\left\{m_{22}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*][\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}_0]\right\}\\
&\qquad+\mathbb{E}_{\theta_0}\left\{O(1)\left(\sum_{j=1}^{n_i}\sum_{s=1}^{S}I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left[\sum_{k=0}^{K_{ij}}\left\{\widehat{\Lambda}_s(Y_{ijk})-\Lambda_{s0}(Y_{ijk})\right\}^2\right]\right)+O(1)\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2\right\}\\
&\quad=\mathbb{E}_{\theta_0}\left\{m_{21}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\mathbb{E}_{\theta_0}\left\{m_{22}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*][\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}_0]\right\}+O_P\left(M^{-2/3}+\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2\right),
\end{aligned}
$$

where $m_{22}(\boldsymbol{\beta},\boldsymbol{\Lambda})[\boldsymbol{h}^*][\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}_0]$ denotes the derivative of $m_2(\boldsymbol{\beta},\boldsymbol{\Lambda})[\boldsymbol{h}^*]$ along the submodel $d\boldsymbol{\Lambda}_0+\epsilon d(\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}_0)$, which is given by equation (S.12) with $\widetilde{h}_s(u)d\Lambda_s(u)$ replaced by $d(\widehat{\Lambda}_s-\Lambda_{s0})(u)$. Then (S.15) may be written as

$$
\begin{aligned}
&\mathbb{E}_{\theta_0}\left\{m_{11}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right\}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\sqrt{M}\mathbb{E}_{\theta_0}\left\{m_{12}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}_0]\right\}\\
&\qquad=-\sqrt{M}\mathbb{P}_M\left\{m_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\right\}+O_P(\sqrt{M}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2)+O_P(M^{-1/6})+o_P(1)
\end{aligned}
$$

and (S.16) as

$$
\begin{aligned}
&\mathbb{E}_{\theta_0}\left\{m_{21}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+\sqrt{M}\mathbb{E}_{\theta_0}\left\{m_{22}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*][\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}_0]\right\}\\
&\qquad=-\sqrt{M}\mathbb{P}_M\left\{m_2(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}+O_P(\sqrt{M}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2)+O_P(M^{-1/6})+o_P(1).
\end{aligned}
$$

Subtracting the above displays and noting that

$$
\mathbb{E}_{\theta_0}\left\{m_{12}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}_0]-m_{22}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*][\widehat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}_0]\right\}=\boldsymbol{0}
$$

by Condition 7, we find

$$
\begin{aligned}
&\sqrt{M}\mathbb{E}_{\theta_0}\left\{m_{11}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)-m_{21}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\\
&\quad=-\sqrt{M}\mathbb{P}_M\left\{m_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)-m_2(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}+O_P(\sqrt{M}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2)+O_P(M^{-1/6})+o_P(1).
\end{aligned}
$$

From Condition 8 and the consistency results of Theorem 1, it follows that

$$
\sqrt{M}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=-\sqrt{M}\mathbb{E}_{\theta_0}\left\{m_{11}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)-m_{21}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}^{-1}\mathbb{P}_M\left\{m_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)-m_2(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\right\}+o_P(1).
$$

Thus $\sqrt{M}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges to a zero-mean normal random vector with covariance matrix I^* given by (S.14) in Condition 8.

B.5 Quadratic Expansion of the Profile Composite Log-Likelihood Function

Note that the joint maximization problem in (S.4) might naturally be reformulated so that the supremum is taken in two steps,

$$
\begin{aligned}
(S.4): \qquad \widehat{\theta}_{CL} &= \operatorname*{argmax}_{\theta \in \mathcal{B} \times \mathcal{C}} \ell_C(\theta; \boldsymbol{\mathcal{O}}) \\
&= \operatorname*{argmax}_{\boldsymbol{\beta} \in \mathcal{B}} \left\{ \operatorname*{argmax}_{\boldsymbol{\Lambda} \in \mathcal{C}} \ell_C(\boldsymbol{\beta}, \boldsymbol{\Lambda}; \boldsymbol{\mathcal{O}}) \right\},
\end{aligned}
$$

meaning that the maximum composite likelihood estimator may equivalently be found by first maximizing $\ell_C(\theta; \boldsymbol{\mathcal{O}})$ over $\mathcal{C}$ for a fixed value of $\boldsymbol{\beta}$ and then by maximizing over $\mathcal{B}$. In particular, let

$$
\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta}) := \operatorname*{argmax}_{\boldsymbol{\Lambda} \in \mathcal{C}} \ell_C(\boldsymbol{\beta}, \boldsymbol{\Lambda}; \boldsymbol{\mathcal{O}})
$$

and define the *profile composite log-likelihood function* by

$$
p\ell_C(\boldsymbol{\beta}) := \sup_{\boldsymbol{\Lambda} \in \mathcal{C}} \ell_C(\boldsymbol{\beta}, \boldsymbol{\Lambda}; \boldsymbol{\mathcal{O}}) = \sum_{i=1}^{M} m \left\{ \boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta}); \boldsymbol{\mathcal{O}}_i \right\}. \tag{S.17}
$$

Then the maximum composite likelihood estimator for $\boldsymbol{\beta}$ is given by

$$
\widehat{\boldsymbol{\beta}} = \operatorname*{argmax}_{\boldsymbol{\beta} \in \mathcal{B}} p\ell_C(\boldsymbol{\beta}) = \operatorname*{argmax}_{\boldsymbol{\beta} \in \mathcal{B}} \mathbb{P}_M \left[m \left\{ \boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta}) \right\} \right]
$$

and the maximum composite likelihood estimator for $\boldsymbol{\Lambda}$ is given by $\widehat{\boldsymbol{\Lambda}} = \widehat{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\beta}})$.

Theorem 3 below provides the second-order asymptotic expansion of the profile composite log-likelihood function $p\ell_C(\widetilde{\boldsymbol{\beta}})$ about $\boldsymbol{\beta}_0$ for any sequence $\widetilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$ We use this result to motivate our proposed profile composite log-likelihood variance estimator in Section 3.3 of the main text.

THEOREM 3: *Under Conditions 1–8, for any sequence $\widetilde{\boldsymbol{\beta}} \overset{P}{\to} \boldsymbol{\beta}_0$,*

$$\begin{aligned}
p\ell_C(\widetilde{\boldsymbol{\beta}}) &= p\ell_C(\boldsymbol{\beta}_0) \\
&\quad + (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \sum_{i=1}^{M} \{m_1(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_2(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\}\,(\boldsymbol{\mathcal{O}}_i) \qquad \text{(S.18)} \\
&\quad - \frac{1}{2} M (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbb{E}_{\theta_0}\{m_{11}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0) - m_{21}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\}\,(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_P(1 + \sqrt{M}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|)^2.
\end{aligned}$$

Proof. We begin by introducing the following map from a neighborhood of $\boldsymbol{\beta} \in \mathbb{R}^p$ into the parameter set $\mathcal{D} = \mathcal{D}_{1,\infty} \times \cdots \times \mathcal{D}_{S,\infty}$ for $\boldsymbol{\Lambda}$:

$$\boldsymbol{\epsilon} \mapsto \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \begin{pmatrix} \Lambda_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \Lambda_1) = \int_0^{(\cdot)} \{1 + (\boldsymbol{\beta} - \boldsymbol{\epsilon})^\top \boldsymbol{h}_1^*(t)\} d\Lambda_1(t) \\ \vdots \\ \Lambda_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \Lambda_S) = \int_0^{(\cdot)} \{1 + (\boldsymbol{\beta} - \boldsymbol{\epsilon})^\top \boldsymbol{h}_S^*(t)\} d\Lambda_S(t) \end{pmatrix}, \qquad \text{(S.19)}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^p$ and $\boldsymbol{h}^* = (\boldsymbol{h}_1^*, \ldots, \boldsymbol{h}_S^*)$, $\boldsymbol{h}_s^*$ a p-dimensional vector of functions in $\mathcal{L}_2(\mu)$, is the least favorable direction satisfying equation (S.13) in Condition 7. Note that this map satisfies: (i) $\boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) \in \mathcal{D}$ for all $\|\boldsymbol{\epsilon} - \boldsymbol{\beta}\|$ sufficiently small and (ii) $\boldsymbol{\Lambda}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \boldsymbol{\Lambda}$ for any $(\boldsymbol{\beta}, \boldsymbol{\Lambda}) \in \mathcal{B} \times \mathcal{D}$.

Let $\eta(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) := m\{\boldsymbol{\epsilon}, \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \boldsymbol{\Lambda})\}$ under the submodel in (S.19). Then

$$\eta(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) = \sum_{i=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left(\sum_{k=0}^{K_{ij}} \delta_{ijk} \log\left[S\{Y_{ijk}; \boldsymbol{\epsilon}, \Lambda_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \Lambda_s)\} - S\{Y_{ij,k+1}; \boldsymbol{\epsilon}, \Lambda_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \Lambda_s)\}\right] \right),$$

and using simple algebra we find that

$$\dot{\eta}(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) := (\partial/\partial\boldsymbol{\epsilon}) \eta(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) = m_1\{\boldsymbol{\epsilon}, \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \boldsymbol{\Lambda})\} - m_2\{\boldsymbol{\epsilon}, \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \boldsymbol{\Lambda})\}[\boldsymbol{h}^*]$$

and

$$\ddot{\eta}(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) := (\partial/\partial\boldsymbol{\epsilon}) \dot{\eta}(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) = m_{11}\{\boldsymbol{\epsilon}, \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \boldsymbol{\Lambda})\} - m_{21}\{\boldsymbol{\epsilon}, \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}(\boldsymbol{\beta}, \boldsymbol{\Lambda})\}[\boldsymbol{h}^*],$$

where m_1, m_2, m_{11}, and m_{21} have the forms given in equations (S.7), (S.8), (S.9), and (S.10), respectively. We note that the maps $(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \mapsto \dot{\eta}(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda})(\boldsymbol{\mathcal{O}}_i)$ and $(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) \mapsto \ddot{\eta}(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda})(\boldsymbol{\mathcal{O}}_i)$ are continuous at $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)$ P-a.e., and that under $(\boldsymbol{\epsilon}, \boldsymbol{\beta}, \boldsymbol{\Lambda}) = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)$, these reduce to

$$\begin{aligned}\dot\eta(\boldsymbol{\beta}_0,\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) &= m_1\{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\} - m_2\{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\}[\boldsymbol{h}^*]\\ &= m_1(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) - m_2(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*]\\ \ddot\eta(\boldsymbol{\beta}_0,\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) &= m_{11}\{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\} - m_{21}\{\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)\}[\boldsymbol{h}^*]\\ &= m_{11}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0) - m_{21}(\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)[\boldsymbol{h}^*].\end{aligned}$$

Thus by arguments similar to those in Lemma S.1 and Lemma S.2, we have that, for some neighborhood V of $(\boldsymbol{\beta}_0,\boldsymbol{\beta}_0,\boldsymbol{\Lambda}_0)$, the class of functions

$$\mathcal{M}_1 := \{\dot\eta(\boldsymbol{\epsilon},\boldsymbol{\beta},\boldsymbol{\Lambda}) : (\boldsymbol{\epsilon},\boldsymbol{\beta},\boldsymbol{\Lambda}) \in V\}$$

is P-Donsker and the class of functions

$$\mathcal{M}_2 := \{\ddot\eta(\boldsymbol{\epsilon},\boldsymbol{\beta},\boldsymbol{\Lambda}) : (\boldsymbol{\epsilon},\boldsymbol{\beta},\boldsymbol{\Lambda}) \in V\}$$

is P-Glivenko-Cantelli. Finally, the arguments in Lemma S.3 may be used to show more generally that

$$\mathbb{E}_{\theta_0}\left(\sum_{j=1}^{n_i}\sum_{s=1}^{S} I(\boldsymbol{Z}_{ij}=\boldsymbol{z}_s)\left[\sum_{k=0}^{K_{ij}}\left\{\widehat\Lambda_s(Y_{ijk};\widetilde{\boldsymbol{\beta}}) - \Lambda_{s0}(Y_{ijk})\right\}^2\right]\right) = O_p(M^{-2/3}) + O(\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2).$$

for any (possibly random) sequence $\widetilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$, so that $\widehat{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{\Lambda}_0$ for any sequence $\widetilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$; these convergence rate results may also be used in conjunction with Taylor expansions to show that

$$P\left[\dot\eta\left\{\boldsymbol{\beta}_0,\widetilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{\beta}})\right\}\right] = o_P(\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| + M^{-1/2}).$$

Then all conditions of Theorem 1 of Murphy and Van der Vaart (2000) are met, and the quadratic expansion in (S.18) holds by similar arguments.

B.6 Additional Lemmas

Lemma S.1: *Under Conditions 1–6, the class of functions*

$$\mathcal{M} = \left\{ \sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left(\sum_{k=0}^{K_{ij}} \delta_{ijk} \log \left[\exp\left\{ - \int_0^{Y_{ijk}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \right.\right.\right.$$
$$\left.\left.\left. - \exp\left\{ - \int_0^{Y_{ij,k+1}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \right] \right) : \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\Lambda} \in \mathcal{D} \right\}$$

is P*-Glivenko-Cantelli, with* $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_S$ *and* $\mathcal{D}_s = \{\Lambda_s : \Lambda_s$ *is a non-decreasing function with* $\Lambda_s(0) = 0$ *and* $\Lambda_s(\tau) < \infty\}$.

Proof. The result follows using the arguments in Lemma 1 of Zeng *et al.* (2017).

Lemma S.2: *Under Conditions 1–6, the class of functions*

$$\mathcal{M}^* = \left\{ \sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left(\sum_{k=0}^{K_{ij}} \delta_{ijk} \log \left[\exp\left\{ - \int_0^{Y_{ijk}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \right.\right.\right.$$
$$\left.\left.\left. - \exp\left\{ - \int_0^{Y_{ij,k+1}} e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(t)} d\Lambda_s(t) \right\} \right] \right) : \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\Lambda} \in \mathcal{D}^* \right\}$$

is P*-Donsker, with* $\mathcal{D}^* = \mathcal{D}_{1,C} \times \cdots \times \mathcal{D}_{S,C}$ *and* $\mathcal{D}_{s,C} = \{\Lambda_s : \Lambda_s$ *is a non-decreasing function with* $\Lambda_s(0) = 0$ *and* $\Lambda_s(\tau) \leqslant C\}$.

Proof. The result follows using the arguments in Lemma 2 of Zeng *et al.* (2017).

Lemma S.3: *Under Conditions 1–6,*

$$\mathbb{E}_{\theta_0} \left(\sum_{j=1}^{n_i} \sum_{s=1}^{S} I(\boldsymbol{Z}_{ij} = \boldsymbol{z}_s) \left[\sum_{k=0}^{K_{ij}} \left\{ \widehat{\Lambda}_s(Y_{ijk}) - \Lambda_{s0}(Y_{ijk}) \right\}^2 \right] \right) = O_p(M^{-2/3}) + O(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2).$$

Proof. Let

$$\overline{m}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) := \log \left[\frac{\exp\{m(\boldsymbol{\beta}, \boldsymbol{\Lambda})\} + \exp\{m(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\}}{2} \right] = \log \left\{ \frac{\mathcal{L}_C(\boldsymbol{\beta}, \boldsymbol{\Lambda}) + \mathcal{L}_C(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)}{2} \right\},$$

where $\mathcal{L}_C(\boldsymbol{\beta}, \boldsymbol{\Lambda})$ is the composite likelihood function for a single cluster, and let

$$\overline{\mathcal{M}}^* := \{ \overline{m}(\boldsymbol{\beta}, \boldsymbol{\Lambda}) : \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\Lambda} \in \mathcal{D}^* \}.$$

It follows from Lemma S.2 and the preservation of the Donsker property under Lipschitz continuous transformations that $\overline{\mathcal{M}}^*$ is P-Donsker. Noting further that the class of functions $\{e^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}(u)} : \boldsymbol{\beta} \in \mathcal{B}\}$ is a VC class with VC-index V, we may use similar arguments to those in Lemma 2 of Zeng *et al.* (2017) to find that the $L_2(P)$ bracketing integral

$$N_{[]}\{\epsilon, \overline{\mathcal{M}}^*, L_2(P)\} = O\left\{\exp\left(\epsilon^{-2V/(V+2)}\right)\right\} \times C/\epsilon.$$

Letting

$$\varphi(\delta) := \int_0^\delta \sqrt{1 + \log N_{[]}\{\epsilon, \overline{\mathcal{M}}^*, L_2(P)\}} d\epsilon,$$

it then follows that $\varphi(\delta) \leqslant O(\delta^{1/2})$.

Given that $\widehat{\boldsymbol{\Lambda}}$ is consistent for Λ (Theorem 1), there exists some finite constant C such that $\widehat{\Lambda}_s(\tau) \leqslant C$ for $s = 1, \ldots, S$ and for large enough M. Thus $\overline{m}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$ belongs to $\overline{\mathcal{M}}^*$. From Theorem 3.4.4 of Van der Vaart and Wellner (1996), we then have that

$$P\left\{\overline{m}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) - \overline{m}(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\right\} \leqslant -cH^2\left\{(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}), (\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\right\},$$

where c is some positive constant and $H\left\{(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}), (\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\right\}$ is the Hellinger distance on the class of independence composite likelihoods, which we note form a class of valid (albeit misspecified) density functions for $\boldsymbol{\mathcal{O}}_i$:

$$H\left\{(\boldsymbol{\beta}, \boldsymbol{\Lambda}), (\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\right\} = \left[\int \left\{\mathcal{L}_C(\boldsymbol{\beta}, \boldsymbol{\Lambda})^{1/2} - \mathcal{L}_C(\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)^{1/2}\right\}^2 d\mu\right]^{1/2}.$$

The above results, along with the consistency of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$ from Theorem 1 and the observation that $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$ also maximizes $\mathbb{P}_M\overline{m}(\boldsymbol{\beta}, \boldsymbol{\Lambda})$, imply that all conditions in Theorem 3.4.1 of Van der Vaart and Wellner (1996) hold. Thus $H\left\{(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}), (\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\right\} = O_P(r_M^{-1})$ for r_M satisfying $r_M^2\varphi(r_M^{-1}) \leqslant \sqrt{M}$. In particular, we can choose $r_M = O(M^{1/3})$, so that $H\left\{(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}), (\boldsymbol{\beta}_0, \boldsymbol{\Lambda}_0)\right\} = O_P(M^{-1/3})$.

The remainder of the proof follows using the arguments in Lemma 3 of Zeng *et al.* (2017).

Web Appendix C. Additional Simulation Results

C.1 Details on Baseline Hazard Generation According to Harden and Kropko (2019)

To highlight both (i) the robustness of the composite EM procedure and profile composite likelihood variance estimator to the shape of the stratum-specific baseline hazard functions and (ii) the ability of the composite EM algorithm to capture arbitrarily complex forms for these baseline functions, we generated $\lambda_s(t)$, $s = 1, \ldots, S$ according to the random spline method of Harden and Kropko (2019) for the scenario I simulations. Briefly, this method generates flexible forms for the stratum-specific baseline survival functions, $S_s(t) = \exp\{-\int_0^t \lambda_s(v)\mu(dv)\}$, by:

(1) partitioning the support $[0, \Psi]$ of the time to event into $(\kappa + 1)$ non-overlapping intervals based on κ randomly selected knot points, $\{t_l : l = 1, \ldots, \kappa\}$, with $t_0 = 0$ and $t_{\kappa+1} = \Psi$;

(2) generating κ uniform random variables, $\{u_l : l = 1 \ldots, \kappa\}$, and setting $y_0 = 1$, $y_{\kappa+1} = 0$, and $y_l = u_{(\kappa-l+1)}$ for $l = 1, \ldots, \kappa$ and $u_{(l)}$ the lth order statistic of $\boldsymbol{u}$; and

(3) fitting a monotonic cubic spline to the points $\{(t_l, y_l) : l = 0, \ldots, \kappa + 1\}$ to form $S_s(t)$.

The baseline density $f_s(t)$ may then be found by taking the derivative of $S_s(t)$, and the baseline hazard function $\lambda_s(t)$ by taking $f_s(t)/S_s(t)$. Under models (10)–(12) in the main text, the resulting individual survival functions for subject j in cluster i with $Z_{ij} = z_s$ are then

$$(10) \qquad S(t|X_{ij}) = S_s(t)^{\exp(\beta X_{ij})}$$

$$(11) \qquad S(t|X_{ij}) = \exp\left[-\int_0^t \lambda_s(u)\exp\{(\gamma_1 + \gamma_2 \log u)X_{ij}\}du\right]$$

$$(12) \; S\{t|X_{ij}, P_{ij}(t)\} = \begin{cases} S_s(t)^{\exp(\alpha_1 X_{ij})} & \text{if } t < \upsilon_{ij} \\ S_s(\upsilon_{ij})^{\exp(\alpha_1 X_{ij}) - \exp\{\alpha_2 + (\alpha_1+\alpha_3)X_{ij}\}} S_s(t)^{\exp\{\alpha_2 + (\alpha_1+\alpha_3)X_{ij}\}} & \text{if } t \geqslant \upsilon_{ij} \end{cases}$$

for $P_{ij}(t) = I(t \geqslant \upsilon_{ij})$. For the simulation studies in Section 4 of the main text, we generated the stratum-specific baseline hazard functions using $\Psi = 320$ and $\kappa = 8$. To simulate times to event T_{ij} under models (10)–(12)—each of which lacks a tractable closed-form expression for both the individual survival function and its inverse—we applied the inverse probability integral transform method to a piecewise-constant approximation to $S_{ij}(t)$, which we denote here by $S^*_{ij}(t)$, and its generalized inverse, which we define as $S^{*-1}_{ij}(\omega) := \inf\{t : S^*_{ij}(t) \leqslant \omega\}$. The number of piecewise components used in this approximation was on the order of 3×10^6.

C.2 Performance of Point and Interval Estimators Under Increasing M and n^*

Table S.1 presents additional simulation results under two different sets of asymptotics on $n = \sum_{i=1}^{M} n_i$. The first two sets of results in Table S.1, corresponding to scenario I with $(S = 4, M = 200, 20 \leqslant n_i \leqslant 30)$ and scenario II with $(S = 20, M = 20, 500 \leqslant n_i \leqslant 700)$, consider an increased M relative to the results in Table 1 of the main text; the third set of results in Table S.1, corresponding to scenario II with $(S = 15, M = 15, 800 \leqslant n_i \leqslant 1000)$, considers an increased n^*.

[Table 1 about here.]

C.3 Sensitivity of the Profile Composite Likelihood Variance Estimator to the Perturbation Constant

Table S.2 summarizes the performance of the profile composite likelihood variance estimator under numerical differentiation with $h_M = cn^{-1/2}$ for $c \in \{0.1, 1, 10\}$; results are presented for scenario I with $(S = 4, M = 100, 20 \leqslant n_i \leqslant 30)$. Figure S.1 displays the profile composite log-likelihood surface under model (11), in which there is a time-varying covariate effect, for a single simulated dataset under $(S = 4, M = 100, 20 \leqslant n_i \leqslant 30)$ and $\tau = 0.5$.

[Table 2 about here.]

[Figure 1 about here.]

C.4 Baseline Hazard Estimation Under Infrequent and Covariate-Dependent Monitoring

Figures S.2 and S.3 illustrate the performance of the stratum-specific baseline survival estimators under model (15) when $M = 100$ and $S = 4$ (Figure S.2) and when $M = S = 15$ (Figure S.3). Estimation results for models (13) and (14) are similar but not shown.

[Figure 2 about here.]

[Figure 3 about here.]

Web Appendix D. Sensitivity Analysis for the Botswana Combination Prevention Project Results

Table S.3 evaluates the sensitivity of the marginal (matched-pair-unadjusted) analysis of the Botswana Combination Prevention Project to the choice of perturbation direction h_M when conducting numerical differentiation for the profile composite likelihood variance estimator.

[Table 3 about here.]

References

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* **39,** 1–22.

Gao, F., Zeng, D., Couper, D., and Lin, D.Y. (2019). Semiparametric regression analysis of multiple right- and interval-censored events. *Journal of the American Statistical Association* **114,** 1232–1240.

Gao, X. and Song, P.X.-K. (2011). Composite likelihood EM algorithm with applications to multivariate hidden Markov model. *Statistica Sinica* **21,** 165–185.

Kosorok, M.R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* New York: Springer.

Murphy, S.A. and Van der Vaart, A.W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95,** 449–465.

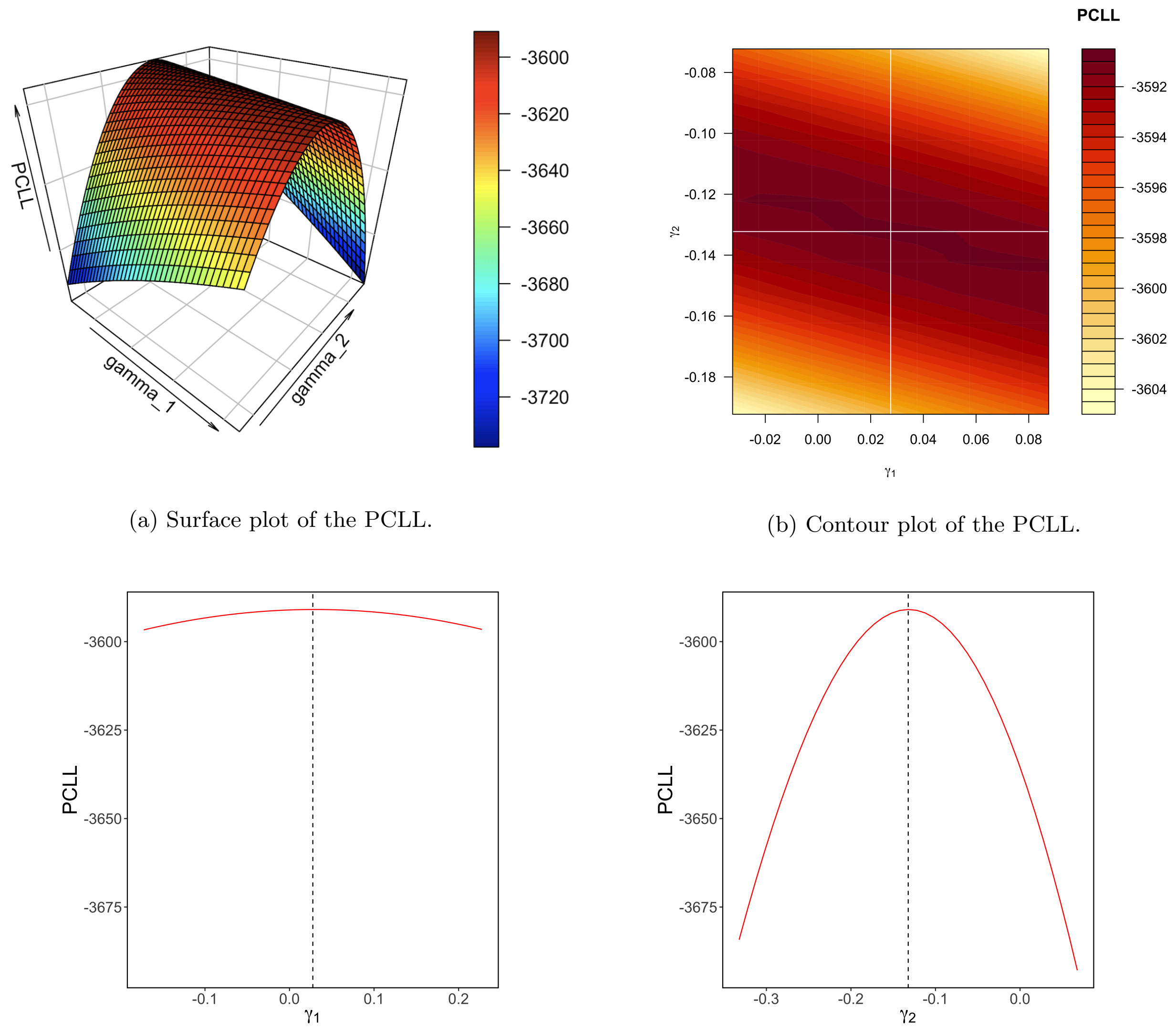
Harden, J.J. and Kropko, J. (2019). Simulating duration data for the Cox model. *Political Science Research and Methods* **7,** 921–928.

Van der Vaart, A.W. (2000). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

Van der Vaart, A.W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* New York: Springer.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11,** 95–103.

Zeng, D., Gao, F., and Lin, D.Y. (2017). Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. *Biometrika* **104,** 505–525.



(a) Surface plot of the PCLL.

(b) Contour plot of the PCLL.

(c) PCLL along $\gamma_2 = \widehat{\gamma}_2$.

(d) PCLL along $\gamma_1 = \widehat{\gamma}_1$.

Figure S.1: Sample profile composite log-likelihood function (PCLL) for model (11) of the main text, in which $\beta(t) = \gamma_1 + \gamma_2 \log t$. The white lines in Figure S.1b denote the maximum composite likelihood estimators for γ_1 and γ_2, which in this dataset were $\widehat{\gamma}_1 = 0.028$ and $\widehat{\gamma}_2 = -0.132$. The maximum PCLL was -3590.926.

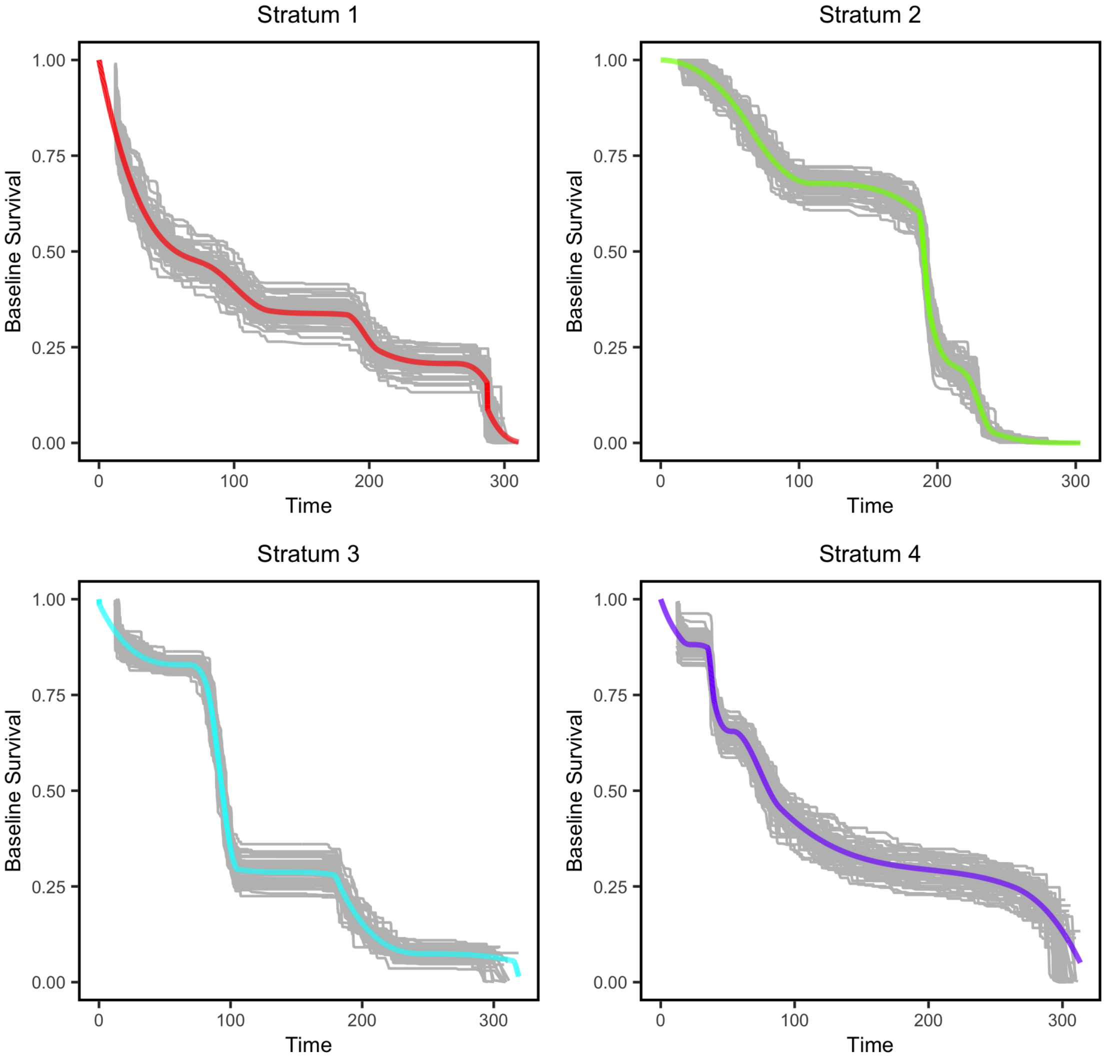


Figure S.2: Comparison of 50 randomly selected estimated stratum-specific baseline survival functions (in gray) with the true data-generating functions (in color) under model (15) with $M = 100$ and $S = 4$.

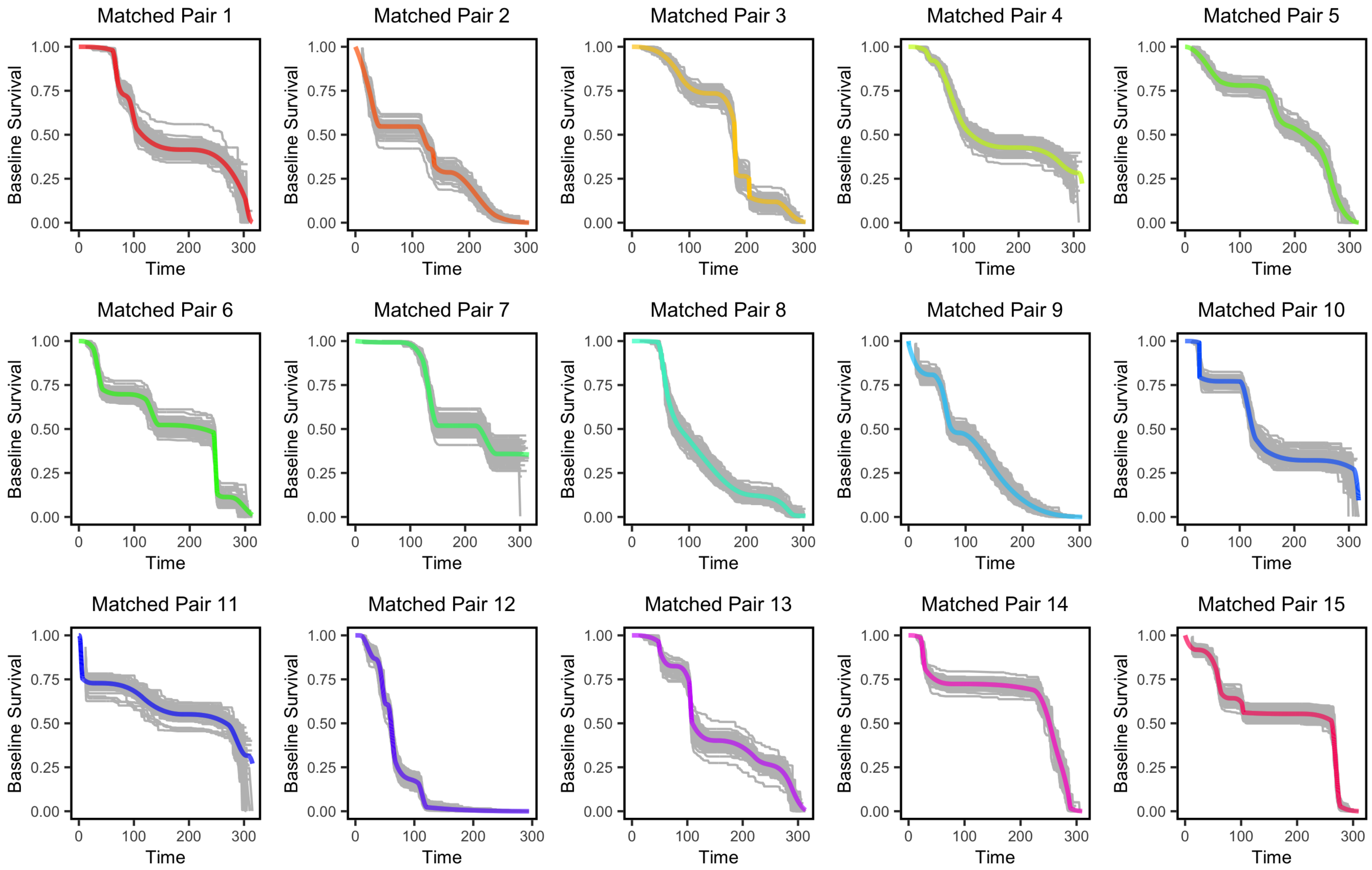


Figure S.3: Comparison of 50 randomly selected estimated stratum-specific baseline survival functions (in gray) with the true data-generating functions (in color) under model (15) with $M = S = 15$.

Table S.1: Finite sample performance of the maximum composite likelihood estimators and profile composite likelihood variance estimators under $h_M = n^{-1/2}$.

	Within-Cluster Independence					Copula Dependence Model				
	Point Est.	Bias	Emp. SE	Est. SE	CP	Point Est.	Bias	Emp. SE	Est. SE	CP
Scenario I: $S = 4$, $M = 200$, $20 \leqslant n_i \leqslant 30$										
Model (10)										
β = -0.30	-0.301	-0.001	0.041	0.042	95.1%	-0.301	-0.001	0.076	0.077	95.0%
Model (11)										
γ_1 = 0.00	0.037	0.037	0.195	0.197	94.2%	0.029	0.029	0.161	0.168	96.3%
γ_2 = -0.15	-0.158	-0.008	0.045	0.045	93.8%	-0.158	-0.008	0.040	0.040	94.7%
Model (12)										
α_1 = -0.30	-0.299	0.001	0.043	0.043	94.9%	-0.303	-0.003	0.079	0.080	95.6%
α_2 = -0.05	-0.051	-0.001	0.057	0.055	94.2%	-0.049	0.001	0.073	0.072	94.5%
α_3 = 0.20	0.203	0.003	0.076	0.074	94.5%	0.195	-0.005	0.098	0.098	95.4%
Scenario II: $S = 20$, $M = 20$, $500 \leqslant n_i \leqslant 700$										
Model (10)										
β = -0.30	-0.300	0.000	0.021	0.021	94.0%	-0.301	-0.001	0.042	0.039	92.3%
Model (11)										
γ_1 = 0.00	0.024	0.024	0.075	0.080	94.6%	0.031	0.031	0.106	0.107	93.2%
γ_2 = -0.15	-0.157	-0.007	0.019	0.022	95.7%	-0.159	-0.009	0.028	0.028	93.4%
Model (12)										
α_1 = -0.30	-0.301	-0.001	0.026	0.024	92.3%	-0.300	0.000	0.043	0.041	92.4%
α_2 = -0.05	-0.050	0.000	0.039	0.037	93.2%	-0.051	-0.001	0.041	0.039	91.8%
α_3 = 0.20	0.201	0.001	0.052	0.049	92.4%	0.200	0.000	0.056	0.054	93.7%
Scenario II: $S = 15$, $M = 15$, $800 \leqslant n_i \leqslant 1000$										
Model (10)										
β = -0.30	-0.301	-0.001	0.021	0.020	92.2%	-0.302	-0.002	0.037	0.036	91.5%
Model (11)										
γ_1 = 0.00	0.018	0.018	0.072	0.075	95.0%	0.025	0.025	0.098	0.099	93.4%
γ_2 = -0.15	-0.155	-0.005	0.019	0.020	94.4%	-0.156	-0.006	0.025	0.026	94.6%
Model (12)										
α_1 = -0.30	-0.302	-0.002	0.024	0.022	91.1%	-0.303	-0.003	0.041	0.038	91.3%
α_2 = -0.05	-0.050	0.000	0.035	0.034	93.6%	-0.051	-0.001	0.038	0.036	90.3%
α_3 = 0.20	0.201	0.001	0.047	0.045	92.0%	0.204	0.004	0.054	0.050	90.5%

Point Est., empirical average of the parameter estimator; Bias, empirical average of the bias; Emp. SE, empirical standard error; Est. SE, empirical average of the standard error estimator; CP, empirical coverage probability of the corresponding 95% Wald-type confidence interval. All results are summarized across 1000 simulation replicates.

Table S.2: Comparison of the profile composite likelihood variance estimator under three choices of perturbation constant for numerical differentiation, $h_M = cn^{-1/2}$. Results shown for scenario I with $M = 100$ under both exchangeable ($\tau = 0$) and hierarchical ($\tau = 0.5$) correlation structures.

				$c = 0.1$		$c = 1$		$c = 10$	
		Point Est.	Emp. SE	Est. SE	CP	Est. SE	CP	Est. SE	CP
Model (10)									
$\beta =$ -0.30	$\tau = 0.0$	-0.303	0.061	0.063	95.5%	0.063	95.3%	0.063	95.7%
	$\tau = 0.5$	-0.305	0.112	0.107	93.4%	0.107	93.3%	0.106	93.2%
Model (11)									
$\gamma_1 =$ 0.00	$\tau = 0.0$	0.034	0.153	0.165	94.4%	0.156	94.8%	0.379	100.0%
	$\tau = 0.5$	0.027	0.285	0.321	93.5%	0.289	95.0%	46.543	100.0%
$\gamma_2 =$ -0.15	$\tau = 0.0$	-0.158	0.040	0.042	94.8%	0.039	94.5%	0.105	100.0%
	$\tau = 0.5$	-0.156	0.063	0.072	94.3%	0.064	95.2%	11.534	100.0%
Model (12)									
$\alpha_1 =$ -0.30	$\tau = 0.0$	-0.301	0.083	0.086	94.3%	0.082	94.1%	0.081	94.3%
	$\tau = 0.5$	-0.304	0.117	0.118	93.7%	0.116	93.7%	0.118	94.3%
$\alpha_2 =$ -0.05	$\tau = 0.0$	-0.049	0.136	0.151	94.3%	0.138	94.6%	0.131	93.0%
	$\tau = 0.5$	-0.049	0.098	0.102	95.5%	0.098	95.5%	0.098	95.6%
$\alpha_3 =$ 0.20	$\tau = 0.0$	0.197	0.175	0.200	95.1%	0.179	95.1%	0.162	93.6%
	$\tau = 0.5$	0.205	0.132	0.141	95.5%	0.136	95.7%	0.126	94.6%

Point Est., empirical average of the parameter estimator; Emp. SE, empirical standard error; Est. SE, empirical average of the standard error estimator; CP, empirical coverage probability of the corresponding 95% Wald-type confidence interval. All results are summarized across 1000 simulation replicates.

Table S.3: Sensitivity of the marginal analysis of the Botswana Combination Prevention Project to the choice of $h_M = cn^{-1/2}$.

		$c = 1$			$c = 5$			$c = 10$		
	$\widehat{\beta}$	Est. SE	95% CI	P-value	Est. SE	95% CI	P-value	Est. SE	95% CI	P-value
Model (13)										
Combination prevention	-0.463	0.194	(-0.844, -0.083)	0.017	0.198	(-0.851, -0.076)	0.019	0.197	(-0.849, -0.077)	0.019
Model (14)										
Combination prevention	-0.463	0.199	(-0.854, -0.072)	0.020	0.198	(-0.851, -0.075)	0.019	0.197	(-0.850, -0.076)	0.019
Universal test-and-treat adoption	0.159	0.445	(-0.714, 1.032)	0.721	0.347	(-0.523, 0.840)	0.647	0.374	(-0.574, 0.892)	0.670
Model (15)										
Combination prevention										
Prior to UTT adoption	-0.138	0.318	(-0.761, 0.485)	0.664	0.323	(-0.771, 0.495)	0.669	0.295	(-0.715, 0.439)	0.639
Post UTT adoption	-0.935	0.401	(-1.722, -0.149)	0.020	0.444	(-1.806, -0.065)	0.035	0.399	(-1.718, -0.152)	0.019
Universal test-and-treat adoption	0.557	0.507	(-0.438, 1.551)	0.273	0.708	(-0.832, 1.945)	0.432	0.537	(-0.496, 1.610)	0.300

Est. SE, estimated standard error for the log hazard ratio; CI, confidence interval for the log hazard ratio.