Biometrics ***, 1–25 DOI: ***

*** ***

Estimation of Conditional Power for Cluster-Randomized Trials with Interval-Censored Endpoints

Kaitlyn Cook[1,*], and Rui Wang[1,2]

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

[2] Department of Population Medicine, Harvard Medical School, Boston, Massachusetts, U.S.A.

*email: kaitlyncook@g.harvard.edu

Summary: Cluster-randomized trials (CRTs) of infectious disease preventions often yield correlated, interval-censored data: dependencies may exist between observations from the same cluster, and event occurrence may be assessed only at intermittent clinic visits. This data structure must be accounted for when conducting interim monitoring and futility assessment for CRTs. In this article, we propose a flexible framework for conditional power estimation when outcomes are correlated and interval-censored. Under the assumption that the survival times follow a shared frailty model, we first characterize the correspondence between the marginal and cluster-conditional survival functions, and then use this relationship to semiparametrically estimate the cluster-specific survival distributions from the available interim data. We incorporate assumptions about changes to the event process over the remainder of the trial—as well as estimates of the dependency among observations in the same cluster—to extend these survival curves through the end of the study. Based on these projected survival functions we generate correlated interval-censored observations, and then calculate the conditional power as the proportion of times (across multiple full-data generation steps) that the null hypothesis of no treatment effect is rejected. We evaluate the performance of the proposed method through extensive simulation studies, and illustrate its use on a large cluster-randomized HIV prevention trial.

Key words: Cluster-randomized trial; Conditional power; Interim monitoring; Interval censoring.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Cluster-randomized trials (CRTs) are well suited for the study of infectious disease prevention and intervention strategies (e.g., Pronyk, et al., 2006; Lemaitre et al., 2009). By collectively randomizing groups of individuals to receive either the intervention or the standard-of-care, CRTs lessen the possibility of treatment contamination across randomization arms and allow investigators to capture both the direct and indirect effects of the intervention (Hayes and Moulton, 2017). However the data generated from these studies are often complexly structured. The "clustering effect" is a well-noted feature of CRTs: two individuals from the same cluster are more likely to be similar to one another than two individuals from different clusters. Furthermore, the outcome of interest in infectious disease CRTs is often a time-to-event outcome, such as time to HIV seroconversion. In the event that this outcome is asymptomatic or otherwise only observable via periodic examination (as is the case for HIV seroconversion), it is also interval-censored. Thus the design, monitoring, and analysis of these CRTs must account for correlated, interval-censored data.

Our focus here lies specifically on the issue of interim monitoring to permit early stopping for efficacy or futility. Interim monitoring refers to the practice of evaluating a trial's progress while the trial is ongoing, and is typically conducted through a series of interim statistical analyses (Proschan, et al., 2006). The results of these analyses guide study decisions regarding sample size re-estimation, resource allocation, and early termination, and thus have important ethical and financial implications.

Two statistical frameworks are commonly used for determining if and when to stop a clinical trial early: group sequential testing (Pocock, 1977; O'Brien and Fleming, 1979; Pampallona and Tsiatis, 1994) and stochastic curtailment (Lan et al., 1982; Lachin, 2005). Group sequential methods calculate the test statistic of interest at each interim look, and stop the trial for either efficacy or futility if this test statistic crosses a pre-determined stopping

boundary (Pampallona and Tsiatis, 1994). Stochastic curtailment approaches, on the other hand, involve calculating the conditional power, the conditional probability of rejecting the null hypothesis of no treatment effect at the scheduled end of the trial, given both the observed interim data and some conjecture about the remainder of the study. The trial stops for futility if this conditional power is low (Lachin, 2005).

Standard conditional power formulae for exponential family outcomes typically rely on the assumption that the test statistic, or some transformation thereof, is an asymptotically Brownian motion with linear drift; under this assumption, the conditional power may be expressed in a straightforward manner in terms of the standard normal cumulative distribution function (Lan et al., 1982; Lan and Wittes, 1988). This framework has since been adopted for the interim monitoring of trials with more complexly-structured outcomes, such as repeated measures or failure time data. In the repeated measures context, for example, Wu and Lan (1992) discussed the conditions under which a test statistic based on the linear mixed effects model could give rise to a discrete Brownian motion. Lin et al. (1999), on the other hand, addressed the monitoring of independent right-censored data, using martingale theory to establish that a broad class of weighted log-rank type statistics, possibly with covariate adjustment, converges to a Gaussian process. Martingale theory is not, however, generally available for interval-censored data due to the difficulty of defining an appropriate filtration. While the history of the associated counting process is a natural choice of filtration in the right-censored setting, this counting process is not well-defined for interval-censored data: when no exact failure times are observed, the value of the counting process at a particular time t, as well as the information generated by that process on the interval $[0, t]$, may be unknown. As such, the formulae used in Lin et al. (1999) cannot be easily extended to interval-censored data, and it is otherwise not apparent that test statistics suitable for analyzing correlated interval-censored data are asymptotically Gaussian

processes. Simulation-based approaches for calculating conditional power with time-to-event endpoints have also been proposed, but these methods are currently limited to independent, right-censored data (Henderson et al., 1991).

Thus, while existing conditional power methods are able to address correlated data and right-censored time-to-event data separately, extensions to data that are both correlated and interval-censored are not available but are much needed. As a case in point, we consider the Botswana Combination Prevention Project (BCPP), a cluster-randomized trial evaluating the impact of combination HIV prevention on the 3-year cumulative incidence of HIV in 30 Botswanan communities (Gaolathe et al., 2016). To measure HIV incidence, a random sample of HIV-negative members from each community was tested annually for HIV, resulting in clustered, interval-censored data. When the study begin in 2013, one component of the combination prevention package was immediate antiretroviral therapy initiation for all HIV-positive individuals with high viral loads; the prevailing standard of care at the time was to initiate treatment only for those individuals with CD4 counts below 350 cells/mm^3. However, during the course of the BCPP, the Botswana Ministry of Health updated the national treatment guidelines for HIV to a universal-test-and-treat approach, recommending that all infected individuals—regardless of CD4 count or viral load level—receive antiretroviral therapy. This change likely reduced the magnitude of the expected intervention effect and raised concerns about the power of the study to detect the updated effect. Interim monitoring of the BCPP for futility was thus particularly salient, but difficult given current interim monitoring and futility analysis methods (Gaolathe et al., 2016).

Motivated by this example, we propose a flexible simulation-based framework for estimating conditional power with correlated, interval-censored data. Our approach models the dependence between outcomes via a cluster-specific frailty and permits calculation of the conditional power under any assumed baseline hazard function and hazard ratio over the

remainder of the study. It may also be used with any final analysis method. The statistical contributions of this paper are thus two-fold: it (i) presents the first conditional power method to explicitly consider correlated interval-censored data and (ii) does so in a manner that permits greater analytical flexibility than traditional formulae-based approaches.

The remainder of this paper is structured as follows. In Section 2, we describe in detail our proposed conditional power method. Section 3 examines the performance of this method across a range of data-generating mechanisms and clustering effects, as well as its sensitivity to misspecification of the dependence and to study design choices such as the number and size of the randomized clusters and the width of the censoring interval. In Section 4 we apply the proposed method to interim data patterned on the BCPP, and in Section 5 we conclude with a brief summary and discussion.

2. Methods

Suppose that we conduct a cluster-randomized trial of M independent clusters (indexed by $i = 1, \ldots, M$), with n_i individuals (indexed by $j = 1, \ldots, n_i$) in cluster i. Each of these individuals is associated with a latent set of K distinct monitoring times, $\{Y_{ijk} : k = 1, \ldots, K\}$, at which the outcome of interest would be assessed, as well as a set of R observation indicators, $\{R_{ijk} : k = 1, \ldots, K\}$, with $R_{ijk} = 1$ if individual j in cluster i is present at the kth inspection. The set of observed monitoring times is then given by $\{Y^*_{ijk} = Y_{ijk}R_{ijk} : k = 1, \ldots, K\}$. Let T_{ij} be the time to event for individual j in cluster i, measured from study entry. Note that, for asymptomatic events, we do not observe this failure time directly, but rather observe the interval $(L_{ij}, U_{ij}]$, where L_{ij} is the last observed monitoring time at which individual j in cluster i tests negative for the presence of the event, and U_{ij} is the first observed monitoring time at which individual j in cluster i tests positive. Individuals who remain event-free at the last observed monitoring time are right-censored with observed interval $(\max_k\{Y^*_{ijk}\}, \infty)$. We assume throughout that the observed monitoring times $\boldsymbol{Y}^*_{ij} = \{Y^*_{ijk} : k = 1, \ldots, K\}$ are

independent of the underlying time to infection, and that the underlying time to infection is independent of the calendar time of study entry.

Suppose that an interim analysis is scheduled for calendar time T_I, and that $T_I^{(i)}$ is the study time of the interim analysis, measured in time since randomization of cluster i. Our proposed conditional power estimation procedure includes the following steps, as detailed in Sections 2.1 to 2.4 below (Figure 1). We (1) estimate the conditional survival function in each of the trial communities through time $T_I^{(i)}$ and then (2) use these estimated curves—as well as assumptions regarding the subsequent event process and underlying dependence structure—to extend the survival functions through the remainder of follow-up. We then (3) use a truncated inverse probability integral transform method to generate complete-trial observations from these extended curves and (4) perform the prespecified final analysis on this complete-trial dataset. Finally, we (5) estimate the conditional power of the trial by repeating this data generation and analysis procedure multiple times, and by calculating the proportion of times that the null hypothesis of no intervention effect is rejected.

[Figure 1 about here.]

2.1 *Estimation of the Interim Survival Functions and Dependence Structure*

Let X_i indicate cluster-level randomization to either intervention ($X_i = 1$) or standard-of-care ($X_i = 0$) at baseline. We assume that individual outcomes are independent conditional on cluster membership, and that the hazard in cluster i can be written as

$$\lambda(t|X_i;\eta_i) = \lambda(t|X_i = 0)\exp(\beta X_i + \eta_i), \tag{1}$$

where $\exp(\eta_i)$ is a cluster-specific frailty and $\lambda(t|X_i = 0)$ is the baseline hazard function when $\eta_i \overset{set}{=} 0$. We also assume that the frailties follow a lognormal distribution, $\exp(\eta_i) \sim \text{LogNormal}(0, \sigma^2)$, as the lognormal model has an appealing computational connection to generalized linear mixed models with random intercepts (Ripatti and Palmgren, 2000) and

admits an intuitive relationship between σ^2 and the coefficient of variation, k:

$$k = \frac{\sqrt{Var\big(\lambda(t|X_i;\eta_i)|X_i\big)}}{\mathbb{E}\big(\lambda(t|X_i;\eta_i)|X_i\big)} = \sqrt{e^{\sigma^2}-1}.$$

Our approach may, however, be easily generalized to other common frailty distributions (Web Appendix A). Throughout we will adopt the convention that $\lambda(t|X_i;\eta_i)$ and $S(t|X_i;\eta_i)$ denote the conditional hazard and survival functions within cluster i; that $\overline{\lambda}(t|X_i)$ and $\overline{S}(t|X_i)$ denote the hazard and survival functions marginalized over cluster membership; and that $\lambda(t|X_i)$ and $S(t|X_i)$ denote the conditional hazard and survival functions when $\eta_i \overset{set}{=} 0$.

Under the shared frailty model in (1), the conditional survival function in cluster i is equivalently given by $S(t|X_i;\eta_i) = S(t|X_i)^{\exp(\eta_i)}$. Provided that we are able to estimate both $S(t|X_i)$ and $\exp(\eta_i)$, we may then take $\widehat{S}(t|X_i;\eta_i) = \widehat{S}(t|X_i)^{\exp(\widehat{\eta}_i)}$ for t in $[0, T_I^{(i)}]$. Estimating the conditional survival functions in this fashion presents several notable advantages over stratified estimation of $S(t|X_i;\eta_i)$ within each cluster: it (i) permits explicit characterization of the underlying dependence structure while (ii) effectively leveraging information from all M clusters. This latter point is of particular importance when the outcome of interest is rare and the amount of information available in cluster i is otherwise small. Thus estimation of $S(t|X_i;\eta_i)$ under model (1) reduces to estimation of the common conditional survival function when $\eta_i = 0$, $S(t|X_i)$, as well as the M cluster-specific frailty terms, $\exp(\eta_i)$.

Methods for the semiparametric estimation of $S(t|X_i)$ according to (1) are, however, generally limited, while methods for the nonparametric estimation of the marginal survival function, $\overline{S}(t|X_i)$, are readily available in standard statistical software (e.g., Turnbull, 1976; Wellner and Zhan, 1997). The relationship between these two curves depends on the assumed frailty distribution (see Web Appendix B for further discussion), but may—under the lognormal frailty model—be approximated by

$$\overline{S}(t|X_i) \approx S(t|X_i)\left[1 + \frac{\sigma^2}{2}\log S(t|X_i)\left\{\log S(t|X_i) + 1\right\}\right]. \tag{2}$$

See Appendix A for the derivation of (2). A semiparametric estimator of $S(t|X_i)$ is then given by the root of $g(x) = x\{1 + \frac{\sigma^2}{2}\log x(\log x + 1)\} - \overline{S}(t|X_i)$, which in turn requires reasonable estimates of $\overline{S}(t|X_i)$ and σ^2.

To that end, we nonparametrically estimate the marginal survival function in each intervention arm according to an independence data likelihood:

$$\prod_{i=1}^{M}\prod_{j=1}^{n_i}\left\{\overline{S}(L_{ij}|X_i) - \overline{S}(U_{ij}|X_i)\right\}.$$

Although the independence likelihood is a misspecified version of the joint distribution—thus precluding valid inference on, for example, second-order terms—it still permits consistent estimation of the marginal parameters $\overline{S}(t|X_i = 0)$ and $\overline{S}(t|X_i = 1)$ (Chandler and Bate, 2007). Under nonparametric estimation, these survival functions are identifiable only up to the equivalence class of right-continuous, non-increasing functions defined by the values of $\{\overline{S}(L_{ij}) : i = 1, \ldots, M; j = 1, \ldots, n_i\} \cup \{\overline{S}(U_{ij}) : i = 1, \ldots, M; j = 1, \ldots, n_i\}$. While convention typically takes the nonparametric maximum likelihood estimator (NPMLE) to be the step function within this equivalence class, we instead use linear interpolation to identify the marginal survival functions wherever the NPMLE is non-unique. This choice is made in recognition of the fact that—in settings where the censoring intervals are wide—the NPMLE may be unidentified on large regions of the support of T_{ij}; it may not be reasonable to assume that the survival function is flat during these inter-monitoring periods.

Recovering $S(t|X_i)$ from $\overline{S}(t|X_i)$ also requires estimates of the variance term σ^2, while final estimation of $S(t|X_i; \eta_i)$ requires estimates of the frailty terms $\boldsymbol{\eta} = \{\eta_i : i = 1, \ldots, M\}$. Due to a lack of reliable software for frailty estimation with correlated interval-censored data, we first transform the interval-censored observations into right-censored data via mid-point imputation (Law and Brookmeyer, 1992) and then fit model (1) using a penalized partial likelihood approach (Therneau et al., 2003). This allows us to iteratively estimate both $\widehat{\sigma}^2$ and $\widehat{\boldsymbol{\eta}} = \{\widehat{\eta}_i : i = 1, \ldots, M\}$. As an aside, we note that—while mid-point imputation has

been shown to produce biased coefficient estimates in Cox models fit to independent data—the simulation studies in Section 4 suggest that any bias in the estimation of σ^2 and $\boldsymbol{\eta}$ has limited impact on the final conditional power estimates (Law and Brookmeyer, 1992).

Given $\widehat{\sigma}^2$ and $\widehat{\boldsymbol{\eta}}$, we take $\widehat{S}(t|X_i)$ to be the root of $g(x) = x\{1+\frac{\widehat{\sigma}^2}{2}\log x(\log x+1)\}-\widehat{\widehat{S}}(t|X_i)$, and may then estimate $\widehat{S}(t|X_i;\eta_i) = \widehat{S}(t|X_i)^{\exp(\widehat{\eta}_i)}$.

2.2 *Specification of the Event Process over the Remainder of the Trail*

In order to calculate the conditional power, we also require some set of assumptions about the shape of the two conditional hazard functions when $\eta_i \overset{set}{=} 0$, $\lambda(t|X_i = 0)$ and $\lambda(t|X_i = 1)$, over the remainder of the trial. Denote these hypothesized future values by $\widetilde{\lambda}(t|X_i)$. Then under the assumption that there are no temporal trends in the event process of interest—so that the survival function $S(t|X_i)$ is the same for all clusters with intervention assignment X_i, regardless of the time of study entry—the complete-trial survival function for the ith cluster, $\widetilde{S}(t|X_i;\widehat{\eta}_i)$ is given by

$$\begin{cases} \widehat{S}(t|X_i)^{\exp(\widehat{\eta}_i)}, & t \leqslant T_I^{(i)} \\ \widehat{S}(T_I^{(i)}|X_i)^{\exp(\widehat{\eta}_i)} \exp\left\{-\int_{T_I^{(i)}}^{t} \widetilde{\lambda}(u|X_i)\exp(\widehat{\eta}_i)du\right\}, & t > T_I^{(i)}. \end{cases} \tag{3}$$

The shape of these projected conditional hazard functions may be informed by scientific knowledge or determined by the investigator: our method permits any reasonable specification of $\widetilde{\lambda}(t|X_i = 0)$ and $\widetilde{\lambda}(t|X_i = 1)$. The projections may also be informed by the available interim data, and we detail one such approach to specifying $\widetilde{\lambda}(t|X_i)$ below.

Within each trial arm, we approximate the conditional hazard function prior to interim via a step function with prespecified and equally-spaced knot points $\{\xi_t : t = 0,\ldots,n\}$, where $\xi_0 = 0$ and $\xi_n = T_I$. Let $\lambda_{X_i(t)}$ be the piecewise component on the interval $[\xi_{t-1},\xi_t)$ in the arm with intervention assignment X_i, so that

$$S(\xi_t|X_i) = \exp\{-\lambda_{X_i(t)}(\xi_t - \xi_{t-1})\}S(\xi_{t-1}|X_i)$$

and

$$\lambda_{X_i(t)} = (\xi_t - \xi_{t-1})^{-1} \left\{ \log S(\xi_{t-1}|X_i) - \log S(\xi_t|X_i) \right\}.$$

We then estimate $\widehat{\lambda}_{X_i(t)}$ by replacing $S(\xi_{t-1}|X_i)$ and $S(\xi_t|X_i)$ with the estimated values of the conditional survival functions when $\eta_i = 0$.

Let Δ_{λ_0} and Δ_{λ_1} be projected multiplicative changes to the conditional hazard function in the standard-of-care arm and in the intervention arm, respectively, over the remainder of the trial. We introduce these parameters to accommodate clinical settings in which the future event process is expected to differ systematically from that observed at interim, perhaps in light of delays in intervention roll-out or mid-trial modifications to the intervention. In the absence of additional information regarding the form of $\widetilde{\lambda}(t|X_i)$, we make the final simplifying assumption that both $\widetilde{\lambda}(t|X_i = 0)$ and $\widetilde{\lambda}(t|X_i = 1)$ are constant, and then set $\widetilde{\lambda}(t|X_i = 0) = \overline{\lambda}_0 \cdot \Delta_{\lambda_0}$ and $\widetilde{\lambda}(t|X_i = 1) = \overline{\lambda}_1 \cdot \Delta_{\lambda_1}$, where $\overline{\lambda}_{X_i} = n^{-1} \sum_{t=1}^{n} \widehat{\lambda}_{X_i(t)}$.

2.3 *Simulation of the End-of-Trial Data*

In order to generate a final complete-trial dataset that captures both the observed interim data and the projected future event process, we must first identify the subset of study participants who are still at risk for the event of interest at interim. To that end, we note that, at the time of the interim analysis, there are three possible outcomes for individual j in cluster i: either (i) they have been observed to have the event of interest; (ii) they have not had the event of interest but remain under active follow-up; or (iii) they have been lost to follow-up prior to time T_I. Note that if missingness at the monitoring times is assumed to be intermittent rather than monotone, there may be no trial participants with interim outcome (iii). All individuals with interim outcomes (i) and (iii) thus have complete data at the time of the interim analysis; those with interim outcome (ii) remain at risk for the event of interest. For this last subset of individuals, we then simulate completed observations by generating

both an underlying time to event and the subsequent observation process. The final complete-trial dataset then consists of the observed records from individuals with interim outcomes (i) and (iii) and the simulated records from individuals with interim outcome (ii).

2.3.1 *Simulation of the underlying time to event.* Each individual j in cluster i with interim outcome (ii) is effectively right-censored at the time of the interim analysis, in that their observed censoring interval is (L_{ij}, ∞). Thus we wish to generate simulated times to event $\widetilde{T}_{ij}$ according to the complete-trial survival function for cluster i given in (3) and subject to the restriction that $\widetilde{T}_{ij} > L_{ij}$. In this way, the simulated $\widetilde{T}_{ij}$ will be consistent with both the observed interim data and the projected event process over the remainder of the trial. To do so, we first define the generalized inverse function $\widetilde{S}_i^{-1}(\omega|X_i;\widehat{\eta}_i) = \inf\{t|\widetilde{S}_i(t|X_i;\widehat{\eta}_i) \leqslant \omega\}$ and let $U \sim \text{Uniform}\big(0, \widetilde{S}_i(L_{ij}|X_i;\widehat{\eta}_i)\big)$. Then $\widetilde{T}_{ij} := \widetilde{S}_i^{-1}(U|X_i;\widehat{\eta}_i)$ follows the desired truncated distribution, with $\widetilde{T}_{ij} > L_{ij}$. Appendix B provides formal justification for this truncated inverse probability integral transform method.

2.3.2 *Simulation of the missingness and interval-censoring mechanisms.* We also require some method for simulating the subsequent observation process, which in turn requires generation of both the latent monitoring times, $\widetilde{Y}_{ijk}$, and the observation indicators, $\widetilde{R}_{ijk}$. While we outline some possible modeling choices below, we note that any reasonable specification of the monitoring and missingness processes reflecting the trial experiences is permitted.

In CRTs and other clinical setting, monitoring times are typically planned a priori, though the actual visit dates of the individual study participants will inevitably vary in practice. Suppose that, at the time of the interim analysis, an additional $0 < K'_{ij} < K$ visits are planned for the remainder of the study, and that these visits are scheduled for study times $\tau_{K-K'_{ij}+1}, \ldots, \tau_K$. To incorporate uncertainty around the scheduling of these visits, we simulate the remaining monitoring times $\widetilde{Y}_{ijk} \sim \text{Uniform}(\tau_k - \delta, \tau_k + \delta)$ for some prespecified $\delta > 0$ and for $k = K - K'_{ij} + 1, \ldots, K$. We also set $\widetilde{Y}_{ij,K+1} := \infty$.

Missingness at these monitoring times may be modeled as either intermittent or monotone. In the case of the former, the observation indicators $\{R_{ijk} : k = K - K'_{ij} + 1, \ldots, K\}$ are a sequence of Bernoulli random variables, where the probability of being observed at time k, π^R_{ijk}, optionally depends on the available baseline covariates and the prior history of missingness. π^R_{ijk} may then be estimated using a generalized linear mixed model with individual- and cluster-specific frailties and a logit link. We then simulate $\widetilde{R}_{ijk} \sim \text{Bernoulli}(\widehat{\pi}^R_{ijk})$.

Alternatively, under monotone missingness, the observation indicators for individual j in cluster i represent a coarsening of some latent loss to follow-up time, C_{ij}, in that $R_{ijk} = I(Y_{ijk} \leqslant C_{ij})$. Simulation of the observation indicators $\{R_{ijk} : k = K - K'_{ij} + 1, \ldots, K\}$ thus reduces to the simulation and appropriate transformation of this time. To that end, we first assume that $C_{ij} \sim \text{Exp}(\lambda_C)$, where λ_C is the rate of study attrition. To estimate λ_C from the observed data, we additionally note that the C_{ij} are, in effect, interval-censored. To see this, consider again the three possible outcomes for individual j in cluster i at interim. Under interim outcome (i), the censoring interval for C_{ij} is simply $[U_{ij}, \infty)$, while under interim outcome (ii) the censoring interval is $[L_{ij}, \infty)$. Under interim outcome (iii), let $Y_{ijK''_{ij}}$ be the time of the last observed visit and $\tau_{ij,K''_{ij}+1}$ be the (planned) time of the first missed visit. Then the corresponding censoring interval is $[Y_{ijK''_{ij}}, \tau_{ij,K''_{ij}+1})$. Regardless of the outcome scenario, we denote the observed interval for C_{ij} as $[L^*_{ij}, U^*_{ij})$, and estimate λ_C by maximizing the independence likelihood. We may then use the truncated inverse probability integral transform to simulate $\widetilde{C}_{ij} \sim \text{Exp}(\widehat{\lambda}_C)$ subject to $\widetilde{C}_{ij} > L^*_{ij}$. The set of corresponding observation indicators is simply $\{\widetilde{R}_{ijk} = I(\widetilde{Y}_{ijk} \leqslant \widetilde{C}_{ij}) : k = K - K'_{ij} + 1, \ldots, K\}$.

Regardless of the assumed missingness mechanism, we set the observation indicator for the final visit time $\widetilde{Y}_{ij,K+1}$ to be $\widetilde{R}_{ij,K+1} := 1$. Then the simulated observation process post interim analysis is given by $\{\widetilde{Y}^*_{ijk} = \widetilde{Y}_{ijk}\widetilde{R}_{ijk} : k = K - K'_{ij} + 1, \ldots, K + 1\}$.

2.3.3 *Generation of the final complete-trial observations.* For each study participant with interim outcome (ii), we first simulate both the underlying time-to-event, $\widetilde{T}_{ij}$, and the subsequent observation process, $\{\widetilde{Y}^*_{ijk} : k = K - K'_{ij} + 1, \ldots, K + 1\}$. Let $\widetilde{\boldsymbol{Y}}^*_{ij}$ represent the corresponding full-trial observation process, comprised of both the observed and simulated inspections: $\widetilde{\boldsymbol{Y}}^*_{ij} = \{Y^*_{ijk} : k = 1, \ldots, K - K'_{ij}\} \cup \{\widetilde{Y}^*_{ijk} : k = K - K'_{ij} + 1, \ldots, K + 1\}$. Then the final simulated observation is $(\widetilde{L}_{ij}, \widetilde{U}_{ij}]$, where $\widetilde{L}_{ij} = \max\{\widetilde{Y}^*_{ijk} : \widetilde{Y}^*_{ijk} \in \widetilde{\boldsymbol{Y}}^*_{ij} \text{ and } \widetilde{Y}^*_{ijk} < \widetilde{T}_{ij}\}$ and $\widetilde{U}_{ij} = \min\{\widetilde{Y}^*_{ijk} : \widetilde{Y}^*_{ijk} \in \widetilde{\boldsymbol{Y}}^*_{ij} \text{ and } \widetilde{Y}^*_{ijk} \geqslant \widetilde{T}_{ij}\}$. A diagram further illustrating the process of generating complete-trial observations is provided in Figure 2.

[Figure 2 about here.]

2.4 *Calculation of the Conditional Power*

For a given specification of $\widetilde{\lambda}(t|X_i)$ and of the monitoring and missingness processes, we generate C complete-trial datasets with correlated interval-censored outcomes, and then conduct the prespecified final analysis on each of these simulated datasets; potential analysis methods for correlated interval-censored data include mixed-effects accelerated failure time modeling (e.g., Komárek and Lesaffre, 2007) and randomization-based inference (Wang and De Gruttola, 2017). We then calculate the conditional power as the proportion of the C p-values that are significant at the desired α level.

2.5 *Sensitivity Analyses*

In order to better guide study termination decisions, we recommend that investigators conduct a series of sensitivity analyses for the conditional power. These sensitivity analyses might take the following form. To account for the presence of estimation or investigator uncertainty, the analyst might first construct confidence intervals for each estimated interim parameter (e.g., the conditional hazard functions) or plausible intervals for each projected component of the event process, and then recalculate the conditional power at various values

across this plausible region. This produces a range of conditional power values consistent with the observed data and/or the investigator's uncertainty about the remainder of the trial. Other potential sensitivity analyses include calculating the conditional power under the null trend or under the minimum clinically-meaningful effect. Finally, in order to evaluate the impact of clustering on the final probability of study success, investigators may also wish to calculate the conditional power under an array of plausible coefficients of variation and associated frailty terms.

3. Simulation Study

We conducted a series of simulation studies to examine the performance of our proposed conditional power method, as well as to characterize its robustness to the number and size of clusters, the width of the censoring interval, and the specification of the frailty distribution.

3.1 *Simulation Settings and Evaluation Procedure*

3.1.1 *Data generating mechanism.* We first generated completed CRTs of $M = 30$ pair-matched clusters and $n_i \sim \text{Uniform}(250, 350)$ individuals in each cluster, with one member of each cluster pair randomized to the intervention. The true time to event for each participant was generated according to model (1) with $\lambda(t|X_i = 0) := \lambda_0$ and $\eta_i \sim \text{N}(0, \sigma^2)$, and was subject to monitoring at study times $Y_{ijk} \sim \text{Uniform}(\tau_k - 4, \tau_k + 4)$ for $k = 1, \ldots, 4$ and τ_k the kth element of $\{52, 104, 156, 208\}$ weeks. Participant drop out was separately modeled as monotone with $C_{ij} \sim \text{Exp}(0.002)$, corresponding to an overall loss to follow-up rate of approximately 10% per year. We interval-censored each time to event as a function of the simulated observation process, and then created the interim analysis datasets by truncating the observations after the completion of all τ_2 monitoring visits.

3.1.2 *Simulation settings and structure.* We considered two possible values for the conditional baseline hazard, $\lambda_0 = 0.001$ and 0.01, and varied the conditional intervention effect

β from $[-1, 1]$ at 0.1 intervals. We also considered $\sigma^2 = 0$, 0.06, and 0.22, corresponding to approximate coefficients of variation of $k = 0$, 0.25, and 0.5, respectively. Our primary simulation study consisted of 1000 simulation replicates for each combination of the baseline hazard, intervention effect, and log-frailty variance. We also conducted a series of more focused simulations in which we varied the number and size of the clusters, the width of the censoring interval, and the frailty distribution. We took $\Delta_{\lambda_0} = \Delta_{\lambda_1} = 1$ throughout.

3.1.3 *Analysis method.* For each generated interim dataset, we evaluated the conditional power by testing the null hypothesis of no intervention effect across 500 sets of projected complete-trial data using a permutation test (Wang and De Gruttola, 2017). Our test statistic was the sum of the within-pair differences in cumulative incidence, $T^{obs} = \sum_{g=1}^{15} \left(\widehat{\Lambda}_{1g} - \widehat{\Lambda}_{0g}\right)$, where g indexes cluster pairs and $\widehat{\Lambda}_{X_i g}$ is the estimated cumulative incidence in the cluster from pair g with intervention assignment X_i. We constructed the permutation null distribution by randomizing treatment assignment within each matched pair, and then sampled $P = 1000$ permutation test statistics T_p^* in order to approximate the corresponding p-value for T^{obs}: $\frac{1+\sum_{p=1}^{P} I(|T_p^*| \geqslant |T^{obs}|)}{P+1}$. The conditional power was taken to be the proportion of these 500 p-values that reached significance at an $\alpha = 0.05$ level.

3.1.4 *Performance evaluation.* To evaluate the overall performance of our conditional power method, we compared the mean conditional power across the 1000 simulation replicates to the simulated power for that setting. While we note that the conditional and unconditional power of a trial are not directly comparable, the mean of the empirical conditional power distribution estimates the trial power. In particular, if μ_s is the empirical probability measure for the observed interim data, $\mathcal{I}$, across the $s = 1000$ simulation replicates and μ is the true probability measure for $\mathcal{I}$, then

$$\int \mathbb{P}\left(p(\mathcal{D}) \leqslant \alpha \middle| \mathcal{I}\right) d\mu_s \approx \int \mathbb{P}\left(p(\mathcal{D}) \leqslant \alpha \middle| \mathcal{I}\right) d\mu = \mathbb{P}\left(p(\mathcal{D}) \leqslant \alpha\right)$$

for any p-value $p(\mathcal{D})$ derived from the complete-trial data, $\mathcal{D}$. Thus, if our conditional power method performs as designed, we would expect the mean conditional power to approximate the simulated power for each simulation setting. We also calculated the conditional power under the modification (unachievable in practice) in which $\widetilde{\lambda}(t|X_i = 0) := \widetilde{\lambda}_0$, $\widetilde{\lambda}(t|X_i = 1) := \widetilde{\lambda}_1$, σ^2, and $\boldsymbol{\eta}$ were all set to their data-generating values. This allowed us to isolate whether any discrepancies between the mean conditional power and the simulated power were the result of the projection procedure, the specification and estimation of the parameter vector, or both.

To examine our ability to recapitulate key features of the original complete-trial datasets, we also recorded for each simulation replicate: the mean projected number of events in each trial arm; the mean projected person-time in each trial arm; the bias and mean squared error of $\widehat{\sigma}^2_c$ as an estimator of $\widehat{\sigma}^2_{orig}$, where $\widehat{\sigma}^2_c$ is the estimated variance parameter in the cth projected dataset and $\widetilde{\sigma}^2_{orig}$ is the estimated variance parameter in the original dataset; and the analogous bias and mean squared errors of $\overline{\lambda}_{0,c}$ and $\widehat{\lambda}_{1,c}$, calculated as in Section 2.2 with knot points at each week of follow-up.

3.2 *Simulation Results*

The proposed conditional power method successfully captured both general trends in study conditional power across the simulation settings (Figure 3), as well as specific fluctuations in evidence strength across the individual interim datasets (Figure 4; Web Figure S3). When $\widetilde{\lambda}_0$, $\widetilde{\lambda}_1$, σ^2, and $\boldsymbol{\eta}$ were all set to their data-generating values, the mean conditional power closely approximated the simulated study power for all baseline hazard, intervention effect, and dependence settings. This close correspondence confirms that—absent any estimation error in either the event process or the correlation structure—the resulting conditional power estimates reasonably capture (on average) the underlying study futility. When $\widetilde{\lambda}_0$, $\widetilde{\lambda}_1$, σ^2, and

$\boldsymbol{\eta}$ were instead estimated as in Sections 2.1 and 2.2—and as might be done in practice—the mean conditional power continued to match the simulated trial power.

In addition to capturing aggregate trends in study futility, our proposed method produced individual conditional power estimates that correlated in a meaningful way with the significance of the individual completed studies (Figure 4; Web Figure S3). Assuming a futility threshold of 20% conditional power, 96.8% of all trials classified as futile at interim had final p-values greater than $\alpha = 0.05$, while 93.0% of all trials classified as not futile at interim had final p-values below $\alpha = 0.05$; the final correct classification rate over all 126,000 simulated trials was 94.1%. Although the accuracy rate dipped as low as 71.3% in the independent data setting with low incidence and small intervention effect, these classification errors were attributable almost entirely to continuing studies that ultimately proved futile. This elevated type II error rate is a reflection of the modest inflation of the mean conditional power estimates relative to the study power observed in those simulation settings with near-null intervention effects (Figure 3). This inflation is likely the result of a floor effect: even when data are representative of a null intervention effect, the observed conditional hazard functions $\widehat{\lambda}_{X_i(t)}$ will inevitably differ from one another slightly, and the resulting conditional power estimates will reflect this spurious effect.

Across all simulation settings considered, the proposed complete-trial projection procedure also reasonably captured the salient features of the original trial data, with the bias and mean squared error of the estimated log-frailty variance and conditional hazards all near zero (Web Figures S4–S5). That being said, the bias, when present, tended to be slightly but persistently negative: the projected complete-trial datasets had, on average, fewer events in each trial arm than did the original complete-trial datasets, resulting in smaller within-arm incidences and lower correlation (Web Tables S2–S3). This discrepancy diminished to near zero when individuals were monitored for the event of interest on a bimonthly (as opposed to annual)

basis (Web Table S5; Web Figure S6). This suggests that the bias was not the result of our projection framework per se, but rather of imprecisions in the estimation of $\widehat{S}(t|X_i)$ due to the limited information available at interim under the annual inspection schedule.

[Figure 3 about here.]

[Figure 4 about here.]

3.3 *Sensitivity to the Study Design and Misspecification of the Frailty Distribution*

Our proposed method produced reasonable conditional power estimates across several modifications to the CRT study design, including variations in the number and size of randomized clusters (Web Table S4) and the frequency of the event inspections (Web Table S5; Web Figure S6). Its performance was also robust to misspecification of the frailty distribution: when the true frailties were gamma-distributed, the conditional power estimates derived under the misspecified lognormal model were nearly identical to those under the correctly-specified gamma model (Web Table S6).

4. The Botswana Combination Prevention Project

The Botswana Combination Prevention Project (BCPP) was a pair-matched CRT designed to evaluate the impact of combination HIV prevention strategies on the population-level three-year cumulative incidence of HIV (Gaolathe et al., 2016; Makhema et al., 2019). Thirty communities in Botswana were pair-matched on the basis of size, pre-existing health services, population age structure and geographic location, with one community in each pair randomized to receive combination HIV prevention and the other to receive an enhanced standard-of-care. An incidence cohort of HIV-negative individuals identified from a 20% random sample from each community was tested for seroconversion at each of three annual visits; an interim analysis was planned following the completion of all one-year visits.

At the beginning of the study in 2013, the combination prevention package included

scaled-up HIV testing and counseling services, linkage-to-care support, enhanced mother-to-child transmission prevention efforts, male circumcision campaigns, and extension of antiretroviral therapy to those infected individuals with high viral load levels. The enhanced standard-of-care included higher testing coverage (due to the annual testing of the incidence cohort) and improved technical support for data management, but otherwise reflected the contemporary standard practice. However, while the trial was ongoing the national HIV treatment guidelines changed: the Botswana Ministry of Health recommended in 2016 that all HIV-positive patients, regardless of CD4 count or viral load levels, initiate antiretroviral therapy (Gaolathe et al., 2016). This change made the care administered to the control communities more similar to that in the intervention communities, raising concerns that the anticipated intervention effect might be reduced and that the trial might be futile as a result.

Actual interim data from the BCPP remain confidential. As such, we used our proposed method to conduct a futility assessment on a simulated interim dataset, generated by applying an agent-based epidemic model to a dynamic network of simulated sexual partnerships (Wang et al., 2014). Both the sexual network and epidemic models were developed to project the intervention effect during the design phase of the BCPP, and the inputs to these models were calibrated to resemble the actual study conditions (see Web Appendix E for details).

4.1 *Futility Assessment Using the Network-Generated Data*

The simulated BCPP trial followed an incidence cohort of 10,465 individuals across the 30 study communities: 5,225 of these individuals belonged to combination prevention communities, and 5,240 to enhanced standard-of-care communities. By the time of the interim analysis, 35 individuals in the combination prevention clusters and 50 individuals in the enhanced standard-of-care clusters had seroconverted, corresponding to cluster-conditional incidence rates of 0.0045 and 0.0064 events per person-year, respectively (estimated conditional hazard ratio: 0.699). Assuming no change to these trends, the conditional power of the simulated

BCPP was estimated to be 0.648. However, if the observed intervention effect were to diminish by 10% over the remainder of the trial in response to the new Ministry of Health treatment guidelines (reducing the hazard ratio from 0.699 to 0.769), the estimated conditional power would drop to 0.478. Reductions in the underlying baseline hazard of seroconversion, as might result from the mid-trial adoption of universal antiretroviral therapy initiation, also led to a modest drop in the estimated conditional power (Table 1). For a full summary of the estimated conditional power results, as well as information regarding the projected total person-time of follow-up and number of HIV seroconversions at study conclusion across the simulated complete-trial datasets, see Table 1. The full analysis was executed in R on a single core of the Harvard Medical School O2 cluster and took approximately 3 hours to complete, for an average time of 18 minutes per conditional power estimate.

[Table 1 about here.]

5. Discussion

We have proposed a flexible framework for calculating the conditional power of CRTs with interval-censored endpoints. Our approach permits any assumed form for the conditional hazard functions over the remainder of the study, and may be used with any final hypothesis test of interest. This second feature presents a notable advantage over standard conditional power formulae, which are typically specific to a given test statistic or class of test statistics. Extensive simulation studies demonstrated that the proposed method produces reasonable conditional power estimates across an array of data-generating models and a wide range of study design parameters, and that it is robust to misspecification of the frailty distribution.

We observed mild inflation in the conditional power estimates in those settings with low study power, most likely due to a floor effect on the estimation of the intervention effect. However, the futility bound for clinical trials with conditional-power-based interim

monitoring is often set between 10% and 30% (Zhang et al., 2017), with recent examples including the oncology trial LUME-Lung 2, for which the futility threshold was set at 20% (Lesaffre et al., 2017), and a psychiatric trial of maintenance treatment of bipolar I disorder, for which the futility threshold was set at 30% (Mahableshwarkar et al., 2017). Under these more conservative futility thresholds, our conditional power procedure remains able to accurately discriminate between futile and non-futile trials at interim, even in the presence of this mild conditional power inflation.

In order to estimate the cluster-conditional survival functions $S(t|X_i;\eta_i)$, we adopted the lognormal shared frailty model given in (1). Few computational methods permit stable fitting of (1) to large clusters of interval-censored observations, and semiparametric estimation of this model with correlated interval-censored data remains an open area of research. As such, we used mid-point imputation to transform the interval-censored observations into right-censored observations, from which we then estimated the log-frailty variance and individual frailty terms. However, mid-point imputation has been shown to produce biased coefficient estimates in (1), with the magnitude of the bias increasing with the width of the censoring intervals. For this reason, we exploited the relationship between the marginal and conditional survival functions to obtain a semiparametric estimator for $S(t|X_i)$, rather than relying on a mid-point imputed estimate. Further work is needed to develop stable procedures for fitting frailty models directly to correlated interval-censored observations.

The main focus of this article has been on proposing a method for calculating point estimates of the conditional power. Clinical decision-making would be further aided by having some tool to characterize uncertainty in these estimates. One natural choice for estimating this variability is the bootstrap 95% confidence interval. However, there are several complications that arise when performing bootstrap resampling on correlated, interval-censored observations, particularly when the outcome of interest is both rare and interval-

censored. CRTs typically randomize only a small number of communities, so that bootstrap resampling at the cluster level is unlikely to provide a reasonable approximation to the sampling distribution. Resampling at the individual level conditional on cluster membership is more amenable to a small M setting, but may produce clusters in which no event has occurred when the outcome of interest is rare; this in turn limits the types of test statistics that can be computed on the bootstrapped interim data. In light of these challenges, we instead recommend conducting sensitivity analyses to assess the conditional power across a range of projections that are compatible with the observed data; these analyses should provide the additional context and uncertainty quantification needed for clinicians to make appropriate and informed trial termination decisions at the futility boundary.

Acknowledgements

This research was supported by grants F31 AI141030, T32 AI007358, R37 AI51164 and R01 AI136947 from the U.S. National Institute of Allergy and Infectious Diseases (NIAID).

References

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

Gaolathe, T., Wirth, K.E., Holme, M.P., Makhema, J., Moyo, S., Chakalisa, U., et al. (2016). Botswana's progress toward achieving the 2020 UNAIDS 90-90-90 antiretroviral therapy and virological suppression goals: a population-based survey. *The Lancet HIV* **3,** e221–e230.

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

Henderson, H.G., Fisher, S.G., Weber, L., Hammermeister, K.E. and Sethi, G. (1991).

Conditional power for arbitrary survival curves to decide whether to extend a clinical trial. *Controlled Clinical Trials* **12,** 304–313.

Komárek, A. and Lesaffre, E. (2007). Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution. *Statistica Sinica* **17,** 549–569.

Lachin, J. M. (2005). A review of methods for futility stopping based on conditional power. *Statistics in Medicine* **24,** 2747–2764.

Lan, K.K.G., Simon, R. and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics* **C1,** 207–219.

Lan, K.K.G. and Wittes, J. (1988). The B-value: a tool for monitoring data. *Biometrics* **44,** 579–585.

Law, C.G. and Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine* **11,** 1569–1578.

Lemaitre, M., Meret, T., Rothan-Tondeur, M., Belmin, J., Lejonc, J.L., Luquel, L., et al. (2009). Effect of influenza vaccination of nursing home staff on mortality of residents: a cluster-randomized trial. *Journal of the American Geriatrics Society* **57,** 1580–1586.

Lesaffre, E., Edelman, M.J., Hanna, N.H., Park, K., Thatcher, N., Willemsen, S., et al. (2017). Statistical controversies in clinical research: futility analyses in oncology—lessons on potential pitfalls from a randomized controlled trial. *Annals of Oncology* **28,** 1419–1426.

Lin, D.Y., Yao, Q. and Ying, Z. (1999). A general theory on stochastic curtailment for censored survival data. *Journal of the American Statistical Association* **94,** 510–521.

Mahableshwarkar, A.R., Calabrese, J.R., Macek, T.A., Budur, K., Adefuye, A., Dong, X., et al. (2017). Efficacy and safety of sublingual ramelteon as an adjunctive therapy in the maintenance treatment of bipolar I disorder in adults: a phase 3, randomized controlled

trial. *Journal of Affective Disorders* **221,** 275–282.

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.

Proschan, M.A., Lan, K.K.G. and Wittes, J.T. (2006) *Statistical Monitoring of Clinical Trials: A Unified Approach.* New York: Springer.

Pronyk, P.M., Hargreaves, J.R., Kim, J.C., Morison, L.A., Phetla, G., Watts, C., et al. (2006). Effect of a structural intervention for the prevention of intimate-partner violence and HIV in rural South Africa: a cluster randomised trial. *The Lancet* **368,**1973–1983.

O'Brien, P.C. and Fleming T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35,** 549—556.

Pampallona, S. and Tsiatis, A.A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference* **42,** 19–35.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64,** 191—199.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56,** 1016–1022.

Therneau, T.M., Grambsch, P.M. and Pankratz, V.S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics* **12,** 154–175.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* **38,** 290–295.

Wang, R. and De Gruttola, V. (2017). The use of permutation tests for the analysis of parallel and stepped-wedge cluster-randomized trials. *Statistics in Medicine* **36,** 2831–2843.

Wang, R., Goyal, R., Lei, Q., Essex, M., and De Gruttola, V. (2014). Sample size considera-

tions in the design of cluster randomized trials of combination HIV prevention. *Clinical Trials* **11,** 309–318.

Wellner, J.A. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association* **92,** 945–959.

Wu, M.C. and Lan, K.K.G. (1992). Sequential monitoring for comparison of changes in a response variable in clinical studies. *Biometrics* **48,** 765–779.

Zhang, Q., Freidlin, B., Korn, E.L., Halabi, S., Mandrekar, S. and Dignam, J.J. (2017). Comparison of futility monitoring guidelines using completed phase III oncology trials. *Clinical Trials* **14,** 48–58.

Supporting Information

Web Appendices, Tables, and Figures referenced in Sections 2.1, 3.2, 3.3, and 4 are available with this paper at the Biometrics website on Wiley Online Library.

Appendix A: Derivation of Equation (2)

Let $\dot{S}(t|X_i;\eta_i=0)$ and $\ddot{S}(t|X_i;\eta_i=0)$ be the first and second partial derivatives of $S(t|X_i;\eta_i)$ with respect to η_i, evaluated at $\eta_i=0$. Taking the second-order Taylor approximation of $S(t|X_i;\eta_i)$ about $\eta_i=0$, we have

$$\begin{aligned} S(t|X_i;\eta_i) &\approx S(t|X_i;\eta_i=0)+\dot{S}(t|X_i;\eta_i=0)(\eta_i-0)+\frac{1}{2}\ddot{S}(t|X_i;\eta_i=0)(\eta_i^2-0) \\ &= S(t|X_i)-\eta_i\Lambda(t|X_i)S(t|X_i)+\frac{\eta_i^2}{2}\Lambda(t|X_i)S(t|X_i)\{\Lambda(t|X_i)-1\}, \end{aligned}$$

so that marginalizing over the frailty distribution yields

$$\begin{aligned} \overline{S}(t|X_i) &\approx \int_{-\infty}^{\infty}\left[S(t|X_i)-\eta_i\Lambda(t|X_i)S(t|X_i)+\frac{\eta_i^2}{2}\Lambda(t|X_i)S(t|X_i)\{\Lambda(t|X_i)-1\}\right]\phi(\eta_i)d\eta_i \\ &= S(t|X_i)\left[1+\frac{\sigma^2}{2}\Lambda(t|X_i)\{\Lambda(t|X_i)-1\}\right], \end{aligned}$$

where $\phi(\eta_i)$ denotes the $N(0, \sigma^2)$ density function.

Appendix B: Justification of Truncated Inverse Probability Integral Transform

To sample observations from the survival function in (3), subject to the restriction that $\widetilde{T}_{ij}$ must be larger than the right-censoring time L_{ij}, we define the generalized inverse function $\widetilde{S}_i^{-1}(\omega|X_i;\widehat{\eta}_i) = \inf\{t|\widetilde{S}_i(t|X_i;\widehat{\eta}_i) \leqslant \omega\}$. Then the truncated inverse probability integral transform procedure draws $U \sim \text{Uniform}\big(0, \widetilde{S}_i(L_{ij}|X_i;\widehat{\eta}_i)\big)$ and sets $\widetilde{T}_{ij} := \widetilde{S}_i^{-1}(U|X_i;\widehat{\eta}_i)$.

Claim 1: The resulting $\widetilde{T}_{ij}$ constitutes a random sample from the target conditional distribution, $\widetilde{S}_i(t|T_{ij} > L_{ij}, X_i;\widehat{\eta}_i)$.

Proof. Consider the probability space $(\Omega, \mathcal{B}, P)$, where $\Omega = \big(0, \widetilde{S}_i(L_{ij}|X_i;\widehat{\eta}_i)\big)$, $\mathcal{B}$ is the Borel σ-algebra on Ω, and $P(\cdot) = \mu(\cdot)/\mu(\Omega)$ with μ the Lebesgue measure. Define $A_\omega = \{\omega : \widetilde{S}_i^{-1}(\omega|X_i;\widehat{\eta}_i) > t\}$ and $B_\omega = \{\omega : \omega \leqslant \widetilde{S}_i(t|X_i;\widehat{\eta}_i)\}$ for $t \in (L_{ij}, \infty)$ and $\omega \in \big(0, \widetilde{S}_i(L_{ij}|X_i;\widehat{\eta}_i)\big)$. Note that the desired result follows immediately if we are able to show that $P(A_\omega) = P(B_\omega)$, where $P(B_\omega) = \widetilde{S}_i(t|X_i;\widehat{\eta}_i)/\widetilde{S}_i(L_{ij}|X_i;\widehat{\eta}_i) = \widetilde{S}_i(t|T_{ij} > L_{ij}, X_i;\widehat{\eta}_i)$.

[$\Rightarrow$] Consider $\omega^* \in A_\omega$. Then $\widetilde{S}_i^{-1}(\omega^*|X_i;\widehat{\eta}_i) > t$, so that $t \notin \{t : \widetilde{S}_i(t|X_i;\widehat{\eta}_i) \leqslant \omega^*\}$. So it follows that $\widetilde{S}_i(t|X_i;\widehat{\eta}_i) > \omega^*$, which further implies that $\omega^* \in B_\omega$ and $A_\omega \subset B_\omega$.

[$\Leftarrow$] Define $B'_\omega = \{\omega : \omega < \widetilde{S}_i(t|X_i;\widehat{\eta}_i)\}$ and consider $\omega^* \in B'_\omega$. By the right continuity of $\widetilde{S}_i(t|X_i;\widehat{\eta}_i)$, there exists $\delta_{\omega^*} > 0$ so that $\widetilde{S}_i(t|X_i;\widehat{\eta}_i) - \widetilde{S}_i(t+\delta_{\omega^*}|X_i;\widehat{\eta}_i) < \widetilde{S}_i(t|X_i;\widehat{\eta}_i) - \omega^* \implies \widetilde{S}_i(t+\delta_{\omega^*}|X_i;\widehat{\eta}_i) > \omega^*$. Then $t+\delta_{\omega^*} \notin \{t : \widetilde{S}_i(t|X_i;\widehat{\eta}_i) \leqslant \omega^*\}$ and $\widetilde{S}_i^{-1}(\omega^*|X_i;\widehat{\eta}_i) \geqslant t+\delta_{\omega^*} > t$. So $\omega^* \in A_\omega$ and $B'_\omega \subset A_\omega$.

Note that $P(B_\omega) = P(B'_\omega)$, and that $B'_\omega \subset A_\omega \subset B_\omega \implies P(B'_\omega) \leqslant P(A_\omega) \leqslant P(B_\omega)$. So $P(A_\omega) = P(B_\omega)$, as desired.

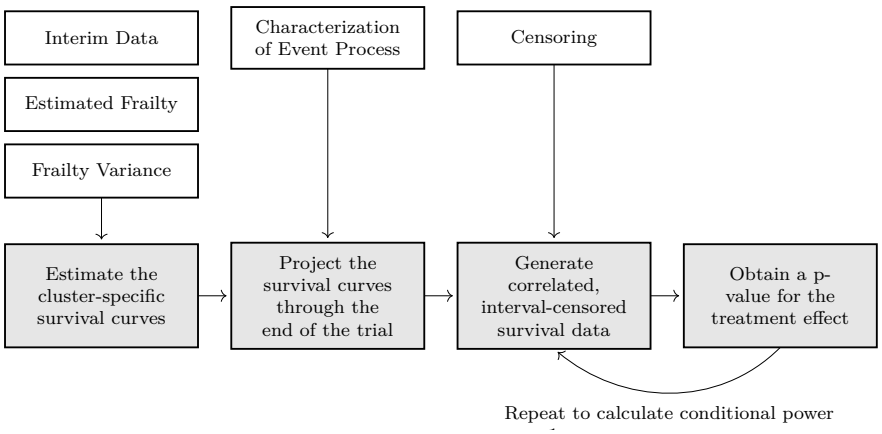

Figure 1. An overview of the proposed conditional power calculation approach. The shaded boxes indicate the analysis pipeline, while the clear boxes indicate inputs to the pipeline that may either be user-specified or informed by the observed interim data.

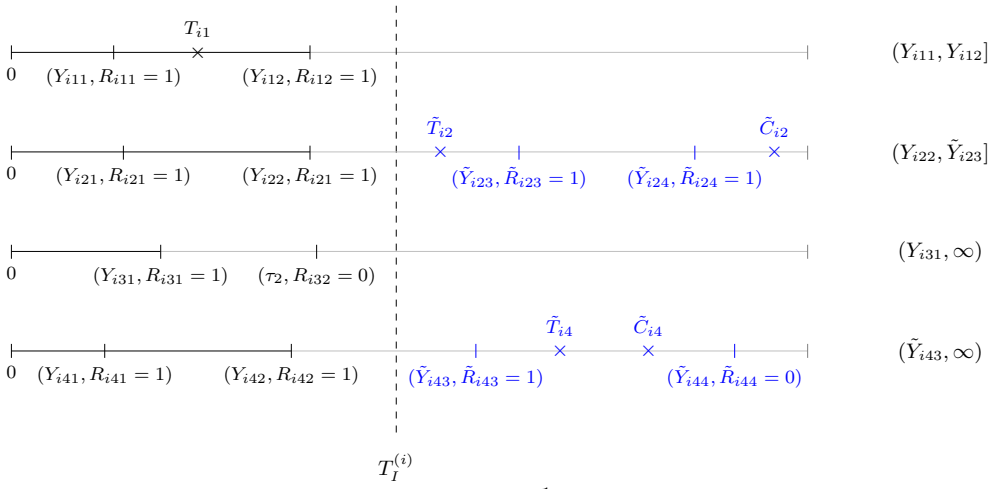

Figure 2. Sample complete-trial generation process under monotone missingness for four members of cluster i, with the observed data rendered in black and the projected data rendered in blue. The final interval-censored observations are given on the right.

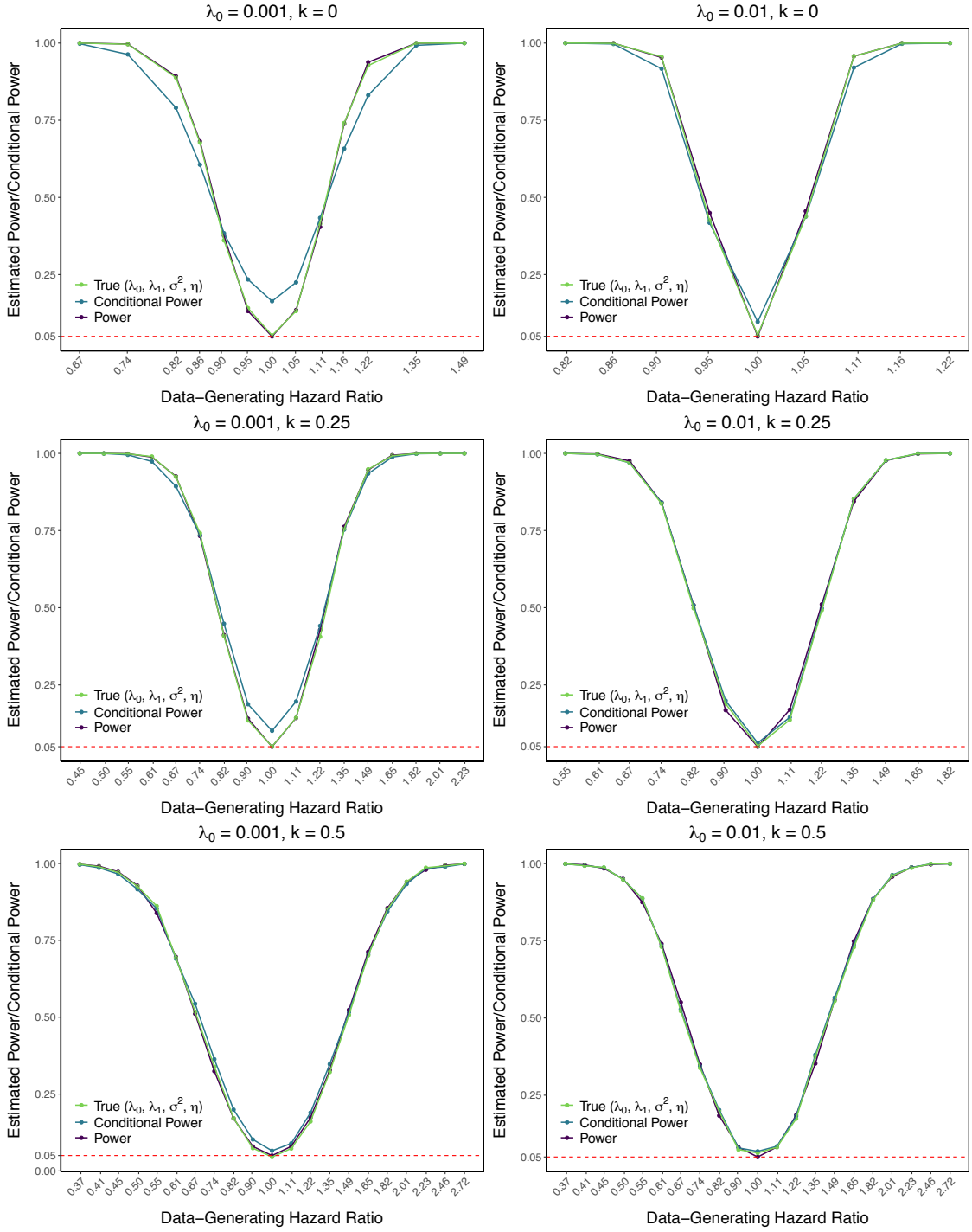

Figure 3. Power and conditional power results as a function of the true, data-generating baseline hazard, intervention effect, and log-frailty variance in the generated interim data. Each data point represents the mean estimated power (conditional power) across 1000 simulation runs.

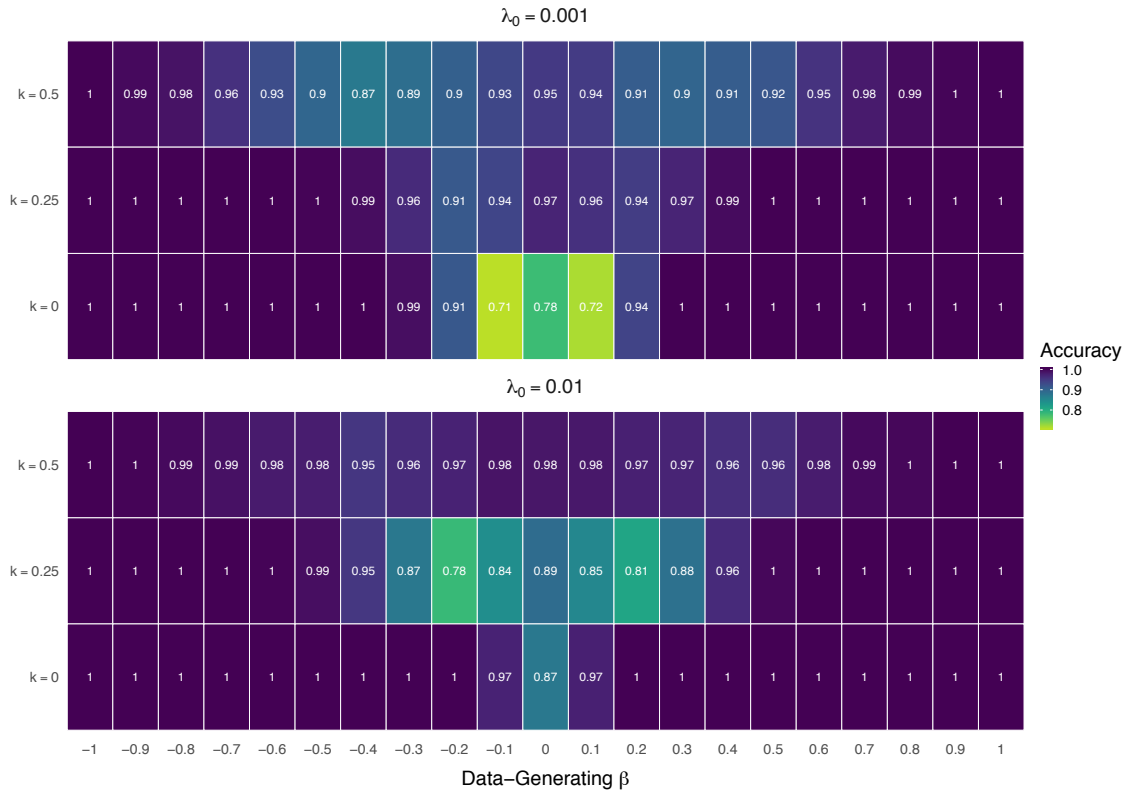

Figure 4. Accuracy rates of the interim futility classifications based on a futility threshold of 20% conditional power and a final significance level of $\alpha = 0.05$.

Table 1

Conditional power and complete-trial projection results for interim data that were patterned on the Botswana Combination Prevention Project (BCPP). These data were generated from an agent-based epidemic model that was spread on a simulated sexual contact network, with inputs informed by BCPP data when available. Number of events, full trial conditional hazard ratios, and person-time are summarized as the mean (and range or standard deviation) across 500 generated end-of-trial datasets; the conditional power is reported as the proportion of those 500 datasets in which the null hypothesis of no effect was rejected.

Δ_{λ_0}	$\widetilde{\lambda}_0$	$\Delta_{\lambda_1}/\Delta_{\lambda_0}$	$\widetilde{\lambda}_1/\widetilde{\lambda}_0$	Number of Events			Person-Time (person-years)		Conditional Power
				Intervention	No Intervention	Full HR	Intervention	No Intervention	
0.9	6.4×10^{-3}	0.8	0.559	56.5 (45, 69)	89.2 (74, 107)	0.632 (0.079)	761,889.4 (1,675.0)	759,082.0 (1,565.8)	0.850
		0.9	0.629	59.5 (46, 75)	88.7 (71, 111)	0.675 (0.088)	761,666.4 (1,592.0)	758,930.1 (1,556.7)	0.696
		1.0	0.699	62.6 (48, 79)	88.7 (71, 110)	0.709 (0.095)	761,432.2 (1,541.9)	759,085.0 (1,649.8)	0.570
		1.1	0.769	65.0 (52, 83)	89.1 (73, 109)	0.737 (0.100)	761,403.3 (1,596.6)	759,148.1 (1,579.1)	0.482
		1.2	0.839	67.7 (54, 89)	88.9 (70, 104)	0.771 (0.108)	761,322.2 (1,586.2)	758,962.1 (1,687.1)	0.356
1.0	5.7×10^{-3}	0.8	0.559	58.9 (45, 73)	93.0 (76, 116)	0.632 (0.085)	761,687.6 (1,656.0)	758,965.1 (1,516.5)	0.856
		0.9	0.629	62.3 (49, 81)	93.0 (75, 117)	0.670 (0.095)	761,502.1 (1,511.4)	758,912.3 (1,585.0)	0.742
		1.0	0.699	64.9 (49, 84)	93.3 (74, 114)	0.697 (0.093)	761,308.8 (1,619.7)	758,875.0 (1,565.2)	0.648
		1.1	0.769	68.0 (54, 86)	93.1 (74, 113)	0.732 (0.097)	761,231.3 (1,651.5)	758,967.4 (1,550.2)	0.478
		1.2	0.839	71.3 (55, 92)	93.2 (75, 117)	0.775 (0.103)	761,294.1 (1,621.0)	758,834.4 (1,656.8)	0.378

Supporting Information for Estimation of Conditional Power for Cluster-Randomized Trials with Interval-Censored Endpoints

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

[1]Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, Massachusetts, U.S.A

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

Web Appendix A. Alternate Frailty Specifications

Suppose that we conduct a cluster-randomized trial of M communities (indexed $i = 1, \ldots, M$) in which the outcome of interest is an interval-censored time-to-event with cluster-specific hazard function

$$\lambda(t|X_i; \eta_i) = \lambda(t|X_i) \exp(\eta_i).$$

While the exposition in the main text assumed $\eta_i \sim \mathrm{N}(0, \sigma^2)$, the proposed method may be generalized to any assumed frailty distribution $f(\exp(\eta); \theta)$ provided that one can (i) estimate θ and $(\eta_1, \ldots, \eta_M)^T$ and (ii) derive either an exact or approximate relationship between the conditional survival function, $S(t|X_i) = S(t|X_i; \eta_i = 0)$, and the induced marginal survival distribution, $\overline{S}(t|X_i)$, in terms of these parameters. Two common alternative choices for the frailty distribution are $\exp(\eta_i) \sim \text{Gamma}(\theta, \theta)$ and $\exp(\eta_i) \sim \text{Positive-Stable}(\alpha)$.

A.1 Gamma Frailty Distribution

Define $\varpi_i := \exp(\eta_i)$, and suppose $\varpi_i \sim \text{Gamma}(\theta, \theta)$ with $\mathbb{E}(\varpi_i) = 1$ and $Var(\varpi_i) = \theta^{-1}$. Then

$$\begin{aligned}
\overline{S}(t|X_i) &= \int_{-\infty}^{\infty} S(t|X_i; \varpi_i) f(\varpi_i; \theta) d\varpi_i \\
&= \int_{-\infty}^{\infty} \exp\left\{-\Lambda(t|X_i)\varpi_i\right\} \frac{\theta^\theta}{\Gamma(\theta)} \varpi_i^{\theta-1} e^{-\theta\varpi_i} d\varpi_i \\
&= \frac{\theta^\theta}{\Gamma(\theta)} \int_{-\infty}^{\infty} \exp\left[-\left\{\Lambda(t|X_i) + \theta\right\}\varpi_i\right] \varpi_i^{\theta-1} d\varpi_i \\
&= \left\{\frac{\theta}{\Lambda(t|X_i) + \theta}\right\}^\theta,
\end{aligned}$$

where $\Lambda(t|X_i) = \int_0^t \lambda(u|X_i)du$ is the cumulative conditional hazard function. Note that θ may be estimated by direct maximization of the marginal likelihood and $(\varpi_1, \ldots, \varpi_M)^T$ by empirical Bayes (see, for example, Klein (1992) and Nielsen et al. (1992)), and that the conditional survival function may be written as an explicit function of $\overline{S}(t|X_i)$ and θ:

$$S(t|X_i) = \exp\left[-\theta\left\{\overline{S}(t|X_i)^{-1/\theta} - 1\right\}\right].$$

Then our proposed conditional power procedure proceeds exactly as described in the main paper, with $\widehat{S}(t|X_i;\widehat{\varpi}_i) = \widehat{S}(t|X_i)^{\widehat{\varpi}_i}$.

A.2 Positive Stable Frailty Distribution

Alternatively, suppose $\varpi_i \sim$ Positive-Stable(α) for $\alpha \in (0,1]$, where the positive stable distributional family is characterized by its Laplace transform

$$\mathbb{E}\{\exp(-c\varpi_i)\} = \exp(-c^\alpha). \tag{A.1}$$

In light of (A.1), we may immediately write down an exact relationship between the marginal and cluster-conditional survival functions, with

$$\overline{S}(t|X_i) = \mathbb{E}\left[\exp\left\{-\Lambda(t|X_i)\varpi_i\right\}\right] = \exp\left\{-\Lambda(t|X_i)^\alpha\right\} = \exp\left[-\left\{-\log S(t|X_i)\right\}^\alpha\right].$$

The frailty parameters α and $(\varpi_1,\ldots,\varpi_M)^T$ may once again be estimated by direct maximization of the marginal likelihood and by empirical Bayes, and the conditional survival function may be estimated using the inverse relationship

$$S(t|X_i) = \exp\left[-\left\{-\log \overline{S}(t|X_i)\right\}^{1/\alpha}\right].$$

Then the remainder of the conditional power procedure proceeds as before, with $\widehat{S}(t|X_i;\widehat{\varpi}_i) = \widehat{S}(t|X_i)^{\widehat{\varpi}_i}$.

Web Appendix B. Marginal & Cluster-Conditional Survival Functions

In this section we discuss the relationship between the conditional and induced marginal survival functions under gamma, positive stable, and lognormal frailty models:

$$\varpi_i \sim \text{Gamma}(\theta,\theta) \qquad \overline{S}(t|X_i) = \left\{\frac{\theta}{\Lambda(t|X_i)+\theta}\right\}^\theta \tag{A.2}$$

$$\varpi_i \sim \text{Positive-Stable}(\alpha) \qquad \overline{S}(t|X_i) = \exp\left\{-\Lambda(t|X_i)^\alpha\right\} \tag{A.3}$$

$$\eta_i \sim \text{N}(0,\sigma^2) \qquad \overline{S}(t|X_i) \approx S(t|X_i)\left[1+\frac{\sigma^2}{2}\Lambda(t|X_i)\{\Lambda(t|X_i)-1\}\right] \tag{A.4}$$

We also comment on the performance of the approximation in (A.4), and demonstrate visually that the approximation error is small in settings with mild to moderate dependence.

B.1 Curve Comparison

(A.2) Under the gamma frailty model, an application of Jensen's inequality demonstrates that the marginal survival function always lies above the conditional curve:

$$\overline{S}(t|X_i) = \mathbb{E}\{S(t|X_i)^{\varpi_i}\} \geqslant S(t|X_i)^{\mathbb{E}(\varpi_i)} = S(t|X_i),$$

with the extent of the discrepancy vanishing as $Var(\varpi_i) = \theta^{-1} \to 0$ or $\Lambda(t|X_i) \to 0$.

(A.3) In the positive stable model, we instead observe that for $\alpha \in (0, 1)$, the marginal survival function lies below the conditional survival function until $\Lambda(t|X_i) = 1$, at which point the two curves cross:

$$\begin{aligned}
\overline{S}(t|X_i) \geqslant S(t|X_i) &\iff \exp\{-\Lambda(t|X_i)^{\alpha}\} \geqslant \exp\{-\Lambda(t|X_i)\} \\
&\iff \Lambda(t|X_i)^{\alpha} \leqslant \Lambda(t|X_i) \\
&\iff \Lambda(t|X_i) \geqslant 1.
\end{aligned}$$

When $\alpha = 1$, the positive stable distribution reduces to a point mass at one, such that all survival times are independent of one another and the marginal and conditional survival functions trivially coincide.

The positive stable distribution also has the notable feature that—assuming the event times follow a Weibull distribution with common shape parameter κ—it preserves proportionality of the hazards under marginalization (Hougaard, 1986). In particular, if the hazard within cluster i is given by

$$\lambda(t|X_i; \varpi_i) = \kappa\lambda^{-1}\,(t/\lambda)^{\kappa-1}\exp(\beta X_i)\varpi_i,$$

then the marginal distribution

$$\overline{S}(t|X_i) = \mathbb{E}\big[\exp\big\{-(t/\lambda)^{\kappa}\exp(\beta X_i)\varpi_i\big\}\big] = \exp\big\{-(t\lambda)^{\alpha\kappa}\exp(\alpha\beta X_i)\big\}$$

is also Weibull, but with both the shape parameter and the log hazard ratio scaled by α; the latter implies an attenuation of the cluster-conditional effect on the marginal scale. Under Weibull failure times the positive stable parameter also has an intuitive interpretation in

terms of the within-cluster correlation: $corr(\log T_{ij}, \log T_{ij'}) = 1 - \alpha^2$ (Hougaard, 1986).

(A.4) In the lognormal frailty model, we see that the marginal survival function will closely approximate the conditional survival function when the contribution of the second term in (A.4) is negligible. This occurs whenever the observations in the study are approximately uncorrelated: either as the result of a small frailty variance, a low study incidence, or minimal observed follow-up time. Furthermore, we see that, as in the positive stable model, $\overline{S}(t|X_i)$ will generally lie below $S(t|X_i)$ whenever $\Lambda(t|X_i) < 1$:

$$\overline{S}(t|X_i) - S(t|X_i) \approx \frac{\sigma^2}{2}\Lambda(t|X_i)\{\Lambda(t|X_i) - 1\}$$

$$\implies sign\big\{\overline{S}(t|X_i) - S(t|X_i)\big\} = sign\big\{\Lambda(t|X_i) - 1\big\}.$$

Thus in our low incidence simulation studies, where $\lambda(t|X_i = 0)$ was set to 0.001, $\overline{S}(t|X_i = 0)$ underestimated $S(t|X_i = 0)$ at all time-points; in our high incidence simulation studies, where $\lambda(t|X_i = 0)$ was set to 0.01, $\overline{S}(t|X_i = 0)$ underestimated $S(t|X_i = 0)$ until $t = 100$ weeks.

B.2 Approximation Performance

As discussed in Section 2.1 of the main text, the approximation in (A.4) permits estimation of the conditional survival function $S(t|X_i)$ as the root of

$$g(y) = y\left\{1 + \frac{\widehat{\sigma}^2}{2}\log y(\log y + 1)\right\} - \widehat{\overline{S}}(t|X_i), \tag{B.1}$$

where $\widehat{\overline{S}}(t|X_i)$ is the nonparametric maximum likelihood estimator of $\overline{S}(t|X_i)$ and $\widehat{\sigma}^2$ is an estimator of the log-frailty variance. Here we consider the extent of the resulting approximation error when both $\overline{S}(t|X_i)$ and σ^2 are known, i.e., when the error is attributable solely to the use of a second-order Taylor approximation; we also compare the performance of $\widehat{S}(t|X_i)$ to the performance of $\overline{S}(t|X_i)$ as (potentially näive) estimators of $S(t|X_i)$. The settings used for this comparison are given in Web Table S1.

[Table 1 about here.]

As anticipated, $\overline{S}(t|X_i)$ underestimated $S(t|X_i)$ when the cumulative incidence was small,

with the extent of this discrepancy increasing in direct correspondence with the log-frailty variance (Web Figures S1 and S2). The Taylor-approximated $\widehat{S}(t|X_i)$ closely matched the true $S(t|X_i)$ in all low incidence settings considered, and diverged substantially from $S(t|X_i)$ only when both the extent of the clustering effect was extreme and the cumulative incidence was large. Note, however, that the performance documented in Web Figures S1 and S2 is an optimistic assessment of $\widehat{S}(t|X_i)$ as an estimator of $S(t|X_i)$ in practice: estimation of both $\overline{S}(t|X_i)$ and σ^2 will necessarily introduce additional error into the approximation. While we anticipate that $\widehat{S}(t|X_i)$ will still provide a better representation of the underlying conditional survival function than $\overline{S}(t|X_i)$, the relative improvement may be small in settings with low dependence.

[Figure 1 about here.]

[Figure 2 about here.]

Web Appendix C. Additional Simulation Results

In this section we provide additional results regarding the ability of our conditional power procedure to:

(C.1) respond to specific fluctuations in evidence strength across the individual study datasets; and

(C.2) generate complete-trial datasets that reasonably capture—in aggregate—the salient features of the observed trial data, had the observed trial run to completion.

To briefly summarize, our simulation study first generated 1000 complete-trial datasets for each unique combination of baseline hazard λ_0, intervention effect β, and log-frailty variance σ^2. We then truncated these datasets following two years of follow-up to create the interim analysis datasets to which our proposed conditional power procedure was applied. Greater

detail regarding the simulation study design and simulation parameters is available in Section 3.1 of the main paper.

C.1 Study-Specific Concordance

Select mosaic and conditional power distribution plots demonstrating our proposed method's study-specific classification performance when $\lambda_0 = 0.001$ and $k = 0.25$ are given in Web Figure S3. The interim futility classification was based on a futility threshold of 20% conditional power, and the final study significance was based on a significance level of $\alpha = 0.05$.

[Figure 3 about here.]

C.2 Generation of Complete-Trial Datasets

Web Figures S4 and S5 presents bias and mean squared error results for the estimated conditional hazards and log-frailty variance in the projected complete-trial datasets, taken as estimators of those same quantities in the original complete-trial data. Web Tables S2 and S3 similarly compare the average number of recorded events, the average number of person-weeks under observation, and the average incidence rates within each arm of the original and projected compete-trial data.

Our projected datasets mildly but persistently underreported the number of events in each trial arm, with this underestimation becoming more prominent in the high incidence setting (Web Figure S5). It is worth noting, however, that the underestimation was largely invariant to the size of the log-frailty variance, and that it persisted even when the data-generating values of λ_0, λ_1, σ^2, and $\boldsymbol{\eta}$ were used to create the projected datasets. In Web Appendix D, we also see that the discrepancy diminished to near zero when individuals were monitored on a more frequent bimonthly inspection schedule. This suggests that the bias noted in Web Figures S4–S5 and the underreporting noted in Web Tables S2–S3 are consequences of the

limited information available at interim under an annual inspection schedule, and not an inherent feature of the method itself.

[Figure 4 about here.]

[Figure 5 about here.]

[Table 2 about here.]

[Table 3 about here.]

Web Appendix D. Sensitivity Analyses

We also conducted a series of focused simulation studies to explore the robustness of our proposed conditional power method to the size and number of randomized clusters, to the width of the censoring interval, and to possible misspecification of the frailty distribution. The simulation design for the sensitivity analyses closely matched that of the primary simulation study (detailed in Section 3.1 of the main paper) except where noted, though we restricted our focus to the low incidence ($\lambda_0 = 0.001$) and low dependence ($k = 0.25$) setting. We also took $\beta \in \{0, -0.2, -0.4\}$, selected to produce trials with approximately 5%, 41.2%, and 92.6% empirical power, respectively. The results of the analyses are presented in Sections D.1–D.3 below.

D.1 Size and Number of Clusters

While many pragmatic CRTs randomize a small number of large clusters—the sort of study design considered in our primary simulation study—studies targeted at a community or household level may randomize a larger number of small clusters (e.g., Guiteras et al. (2015)). In acknowledgment of these alternative CRT designs, we conducted sensitivity analyses under two different large M scenarios: one in which $M = 90$ and $n_i \sim \text{Uniform}(50, 150)$, and the second in which $M = 300$ and $n_i \sim \text{Uniform}(10, 50)$. In both settings, the expected total

sample size, $\sum_i n_i = 9000$, matched that from the primary simulation study. Our results suggest our conditional power projection procedure is robust to the size and number of randomized clusters: it produces reasonable conditional power estimates (Web Table S4) even when $M \approx n_i$ (scenario one) or $M \gg n_i$ (scenario two). Complete-trial data generation and study-specific concordance results are omitted, but were broadly similar to those in Web Appendix C and the main text.

[Table 4 about here.]

D.2 Width of the Censoring Interval

In trials with interval-censored outcomes, the event monitoring schedule that participants follow has potentially large implications for the amount of information available at the interim analysis. The less frequent the inspections, the wider the censoring intervals and the more uncertain the estimates of the within-cluster dependence, the community-specific frailty terms, and the observed trends in incidence. This, in turn, has implications for the quality of the conditional power projections.

Both the BCPP and our primary simulation study operated on one end of the extreme: the inspections followed an annual schedule, and study participants had at most two inspections prior to the interim analysis. To determine the extent to which this monitoring schedule may have contributed to mild under-projection of the number of events (Web Figure S4–S5; Web Tables S2–S3) and mild inflation in the conditional power estimates (Figure 3), we conducted a second sensitivity analysis assuming $K = 26$ bimonthly inspections planned at $\tau_k = 8k$ weeks, $k = 1, \ldots, 26$. As in the primary simulation study, individual participant monitoring times then followed a Uniform$(\tau_k - 4, \tau_k + 4)$ distribution about these planned visits. Finally, for this sensitivity analysis only, we considered both low ($\lambda_0 = 0.001$) and high ($\lambda_0 = 0.01$) incidence settings, as the downward bias noted in the projected number of events appeared to increase with the underlying event rate.

Conditional power estimates under both the low and high incidence settings are given in Web Table S5, while the ability of the conditional power procedure to recapitulate features of the original completed trials when $\lambda_0 = 0.01$ is given in Web Figure S6 (results when $\lambda_0 = 0.001$ were similar to those in Web Appendix C). Crucially, we see that the persistent underestimation of the conditional incidence rates noted in the main simulation study disappeared with more frequent bimonthly inspections, where the bimonthly inspections provided correspondingly more information at interim with which to estimate the conditional survival curves.

[Table 5 about here.]

[Figure 6 about here.]

D.3 Specification of the Frailty Distribution

To examine the robustness of our proposed conditional power method to misspecification of the frailty distribution, we conducted a final sensitivity analysis in which the data-generating model was

$$\lambda(t|X_i;\varpi_i) = 0.001\exp(\beta X_i)\varpi_i,$$

with $\varpi_i \sim \text{Gamma}(1/0.06, 1/0.06)$. We calculated the conditional power under the following three scenarios:

(1) The frailty distribution was misspecified as lognormal, and equation (B.1) was used to estimate the conditional survival functions, $\widehat{S}(t|X_i)$; the projected/estimated $\widetilde{\lambda}_0$, $\widetilde{\lambda}_1$, $\widehat{\sigma}^2$, and $\widehat{\boldsymbol{\eta}}$ were then used to extend these functions through the remainder of the trial.

(2) The frailty distribution was correctly specified, and equation (A.2) was used to estimate the conditional survival functions, $\widehat{S}(t|X_i)$; the projected/estimated $\widetilde{\lambda}_0$, $\widetilde{\lambda}_1$, $\widehat{\theta}$, and $\widehat{\boldsymbol{\varpi}}$ were then used to extend these functions through the remainder of the trial.

(3) The frailty distribution was correctly specified, and equation (A.2) was used to estimate

the conditional survival functions, $\widehat{S}(t|X_i)$; the data-generating λ_0, λ_1, θ, and $\boldsymbol{\varpi}$ were then used to extend these functions through the remainder of the trial.

As shown in Web Table S6, misspecification of the frailty distribution had a negligible impact on the final conditional power estimates: the mean conditional power under scenarios (1) and (2) differed by at most 0.005 points, and the two models produced identical interim futility classifications for 2,979 of the 3,000 generated interim datasets. Complete-trial data generation and study-specific concordance results are omitted, but demonstrated similar trends to those observed in Web Appendix C.

[Table 6 about here.]

Web Appendix E. BCPP Data Generation

We used the same agent-based epidemic model as in Wang et al. (2014), developed during the design stage of the BCPP to project the study-specific incidence of HIV, to generate an interim analysis dataset modeled after the trial. The simulation approach consisted of first generating sexual networks representative of the trial communities before then propagating disease transmission on these networks.

The simulated study-wide sexual network was comprised of 15 independent sub-networks, each corresponding to a matched pair of intervention and standard-of-care communities; these sub-networks captured all heterosexual partnerships between and within the matched communities. For each sub-network, the distribution of number of partnerships per individual was estimated from comprehensive sexual network data from Likoma Island, Malawi (Helleringer and Kohler, 2007), and the extent of sexual mixing between the two matched communities was informed by a pilot study in Mochudi, Botswana (Wang et al., 2014). We used the method of Goyal et al. (2013) to generate sexual networks compatible with these two distributional constraints. The duration of each generated partnership was then drawn

from a survival distribution estimated from data collected in the Mochudi study, and the start date of each generated partnership was drawn uniformly at random over the study period.

We then propagated HIV on these networks using an agent-based epidemic model, in which both individual and community characteristics informed the spread of disease. At the start of the simulation, each individual in the collection of networks was assigned an infection status at random based on current estimates of HIV prevalence in Botswana, reported in Gaolathe et al. (2016) as 29%. Infected individuals were then assigned a viral load category and CD4 count, and characteristics such as transmission probabilities, individual risk-increasing and risk-reducing behaviors, and projected intervention uptake and effects were used to determine HIV spread. A complete listing of the input parameters is given in Web Table S7.

We assumed that a random sample of 20% of each community was selected to receive yearly HIV tests, and that the remainder of the community was tested at a background rate corresponding to the community's intervention or standard-of-care status. The simulation ran for a total of three years; we truncated the data at 82 weeks for the purposes of creating an interim analysis dataset.

[Table 7 about here.]

References

Gaolathe, T., Wirth, K.E., Holme, M.P., Makhema, J., Moyo, S., Chakalisa, U., et al. (2016). Botswana's progress toward achieving the 2020 UNAIDS 90-90-90 antiretroviral therapy and virological suppression goals: a population-based survey. *The Lancet HIV* **3,** e221–e230.

Goyal, R., Blitzstein, J. and De Gruttola, V. (2013). Simulating bipartite networks to reflect

uncertainty in local network properties. *Harvard University Biostatistics Working Paper Series*.

Guiteras, R., Levinsohn, J. and Mobarak, A.M. (2015). Encouraging sanitation investment in the developing world: a cluster-randomized trial. *Science* **348,** 903–906.

Helleringer, S. and Kohler, H.-P. (2007). Sexual network structure and the spread of HIV in Africa: evidence from Likoma Island, Malawi. *AIDS* **21,** 2323–2332.

Hougaard, P. A class of multivariate failure time distributions. *Biometrika* **73,** 671–678.

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48,** 795–806.

Nielsen, G.G., Gill, R.D., Andersen, P.K. and Sørensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19,** 25–43.

Wang, R., Goyal, R., Lei, Q., Essex, M. and De Gruttola, V. (2014) Sample size considerations in the design of cluster randomized trials of combination HIV prevention. *Clinical Trials* **11,** 309–318.

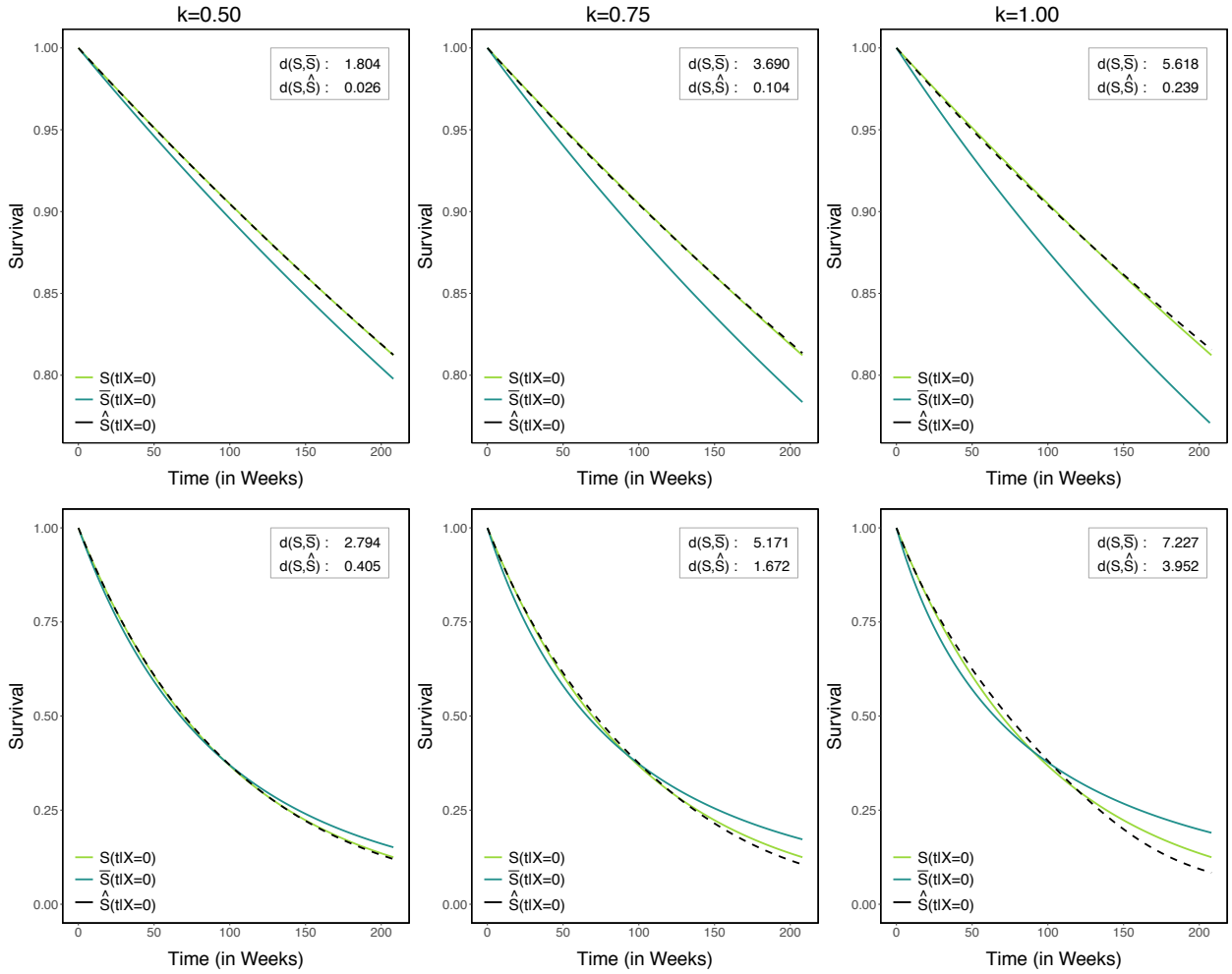

Figure S1. Comparison of the true conditional survival function, $S(t|X_i)$, the induced marginal function, $\overline{S}(t|X_i)$, and the Taylor-approximated conditional function, $\widehat{S}(t|X_i)$, under exponential-distributed failure times and low (top row) and high (bottom row) incidence settings. The distance metrics $d(S, \overline{S}) := \|S - \overline{S}\|_1$ and $d(S, \widehat{S}) := \|S - \widehat{S}\|_1$ were defined over the set $[0, 208]$.

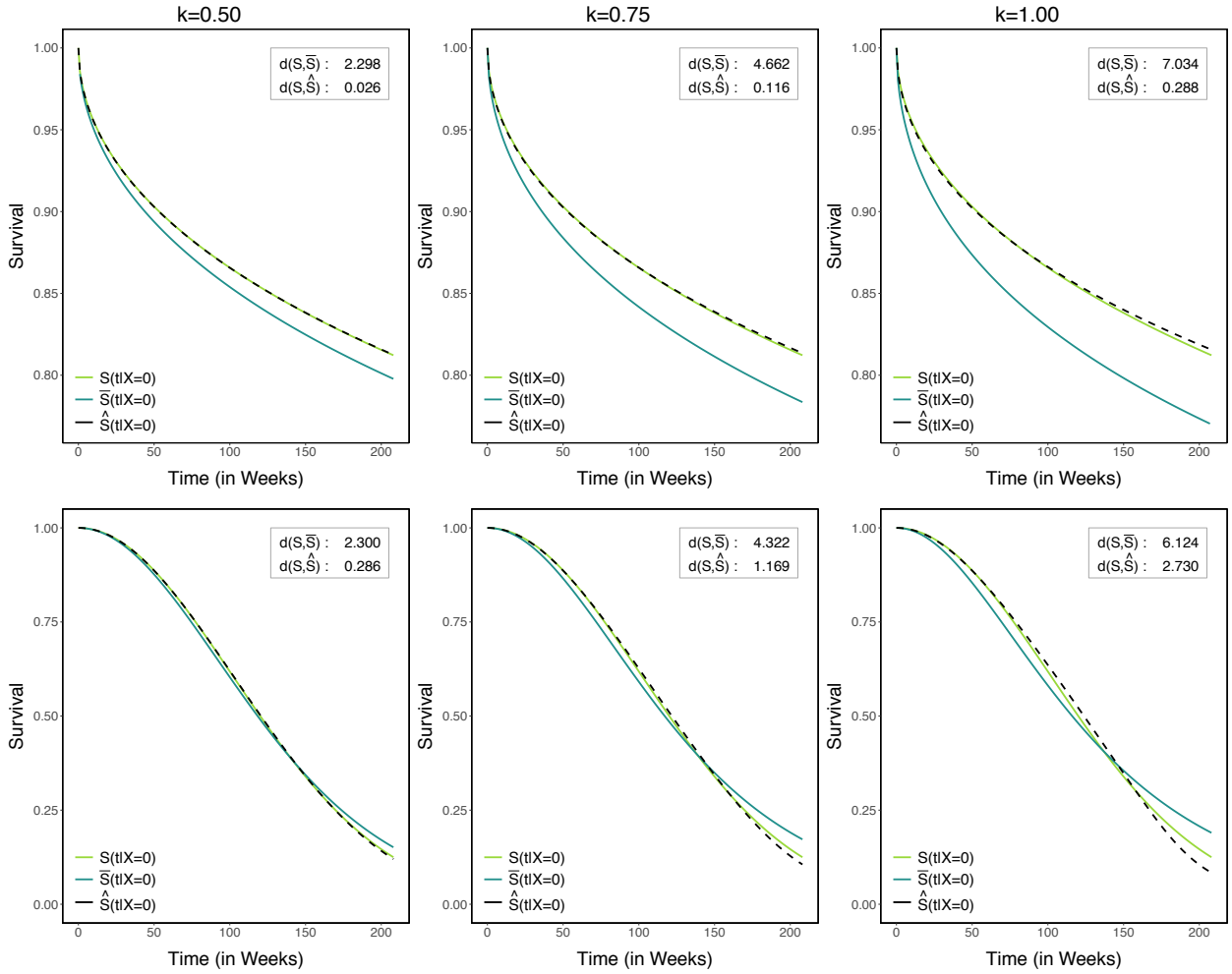

Figure S2. Comparison of the true conditional survival function, $S(t|X_i)$, the induced marginal function, $\overline{S}(t|X_i)$, and the Taylor-approximated conditional function, $\widehat{S}(t|X_i)$, under Weibull-distributed failure times and low (top row) and high (bottom row) incidence settings. The distance metrics $d(S, \overline{S}) := \|S - \overline{S}\|_1$ and $d(S, \widehat{S}) := \|S - \widehat{S}\|_1$ were defined over the set $[0, 208]$.

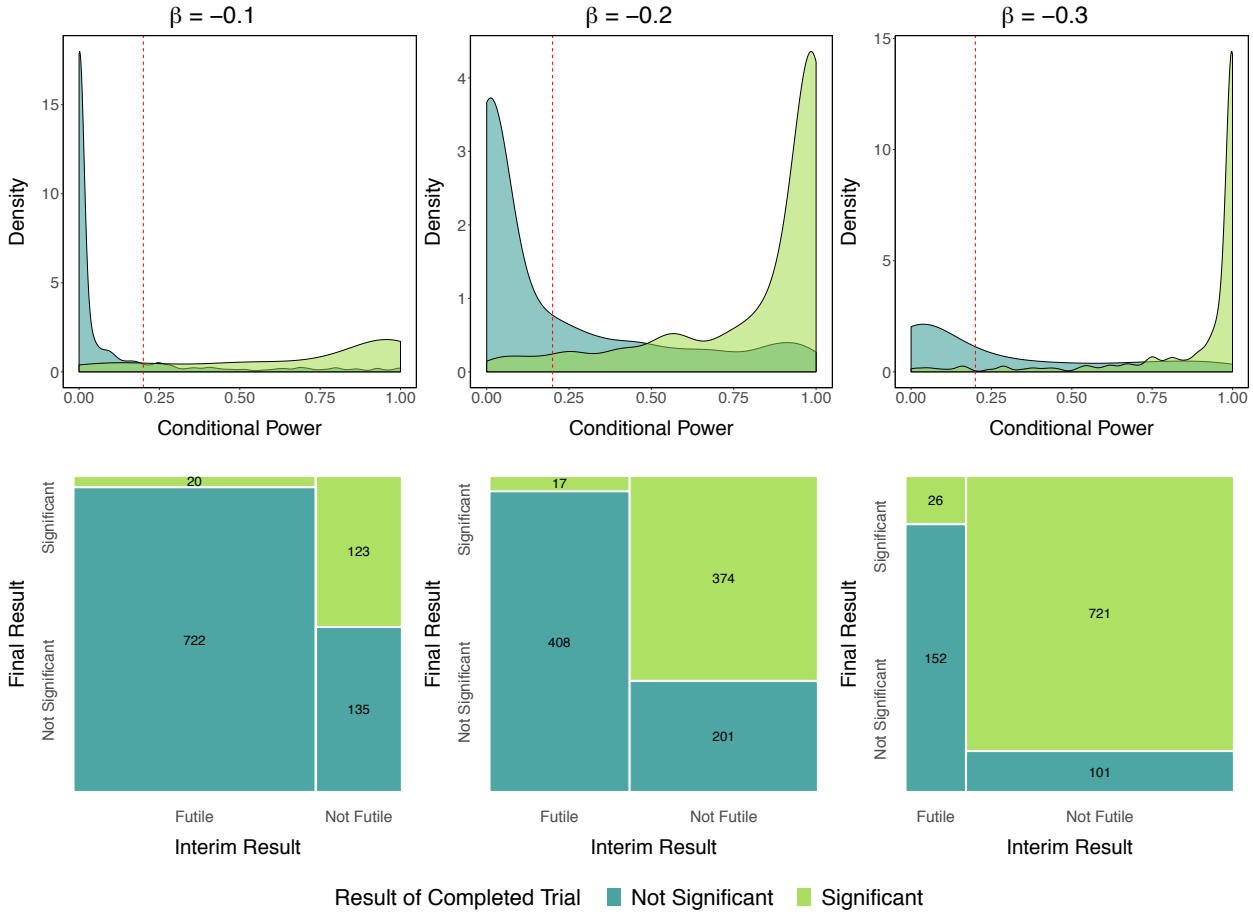

Figure S3. Distribution of interim conditional power estimates stratified by significance at the $\alpha = 0.05$ level of the completed trial (top row), as well as the corresponding classification performance assuming a futility threshold of 20% conditional power (bottom row), for select low incidence ($\lambda_0 = 0.001$) and low dependence ($k = 0.25$) simulation settings. All plots are summarized over 1000 simulated interim datasets.

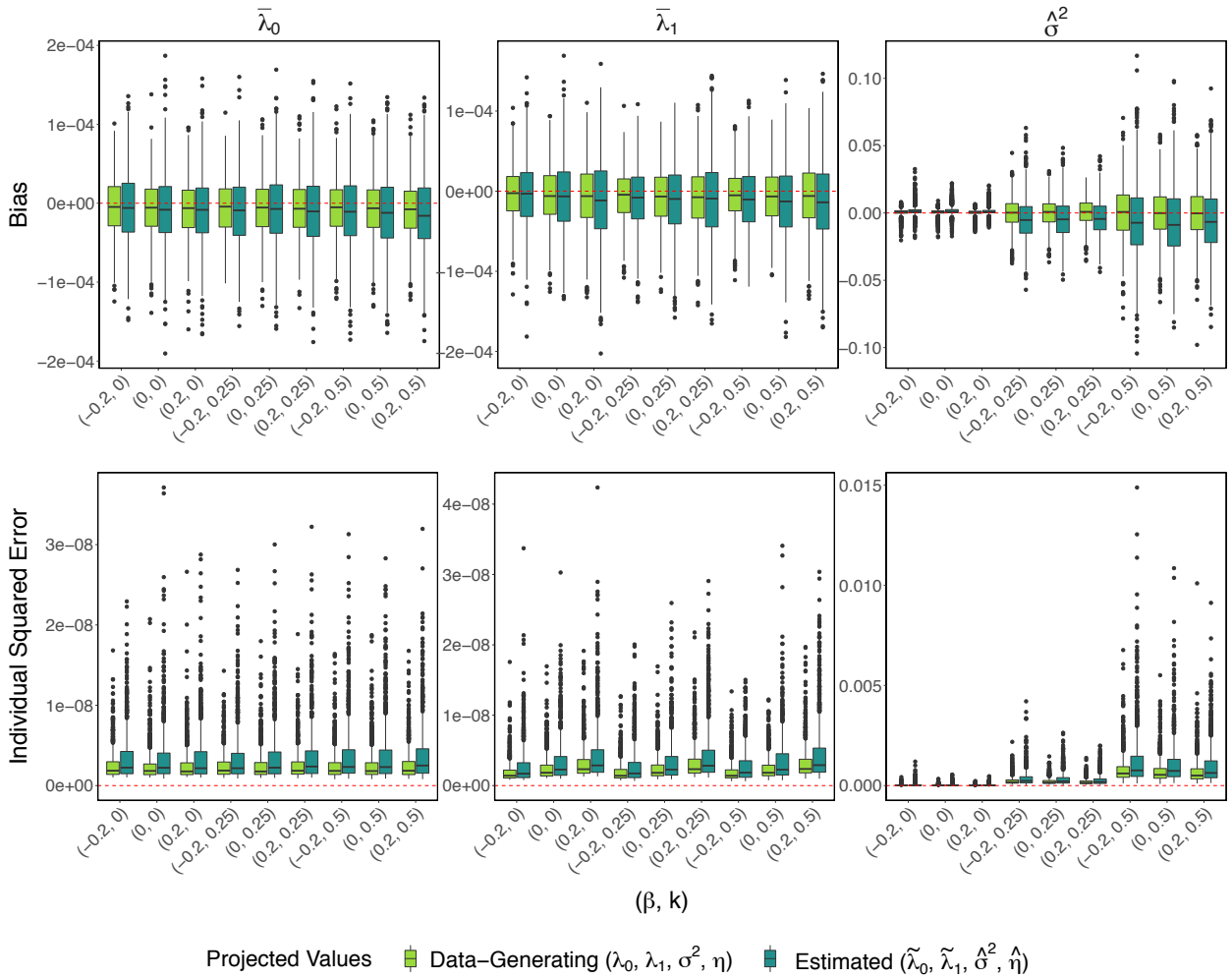

Figure S4. Bias and squared error of the estimated conditional hazards (left, center) and log-frailty variance (right) in the projected complete-trial datasets as estimators of the analogous quantities in the original completed trials. The original data were generated with $\lambda_0 = 0.001$, and selected results are presented for $\beta = (-0.2, 0.0, 0.2)$ and $k = (0.00, 0.25, 0.50)$.

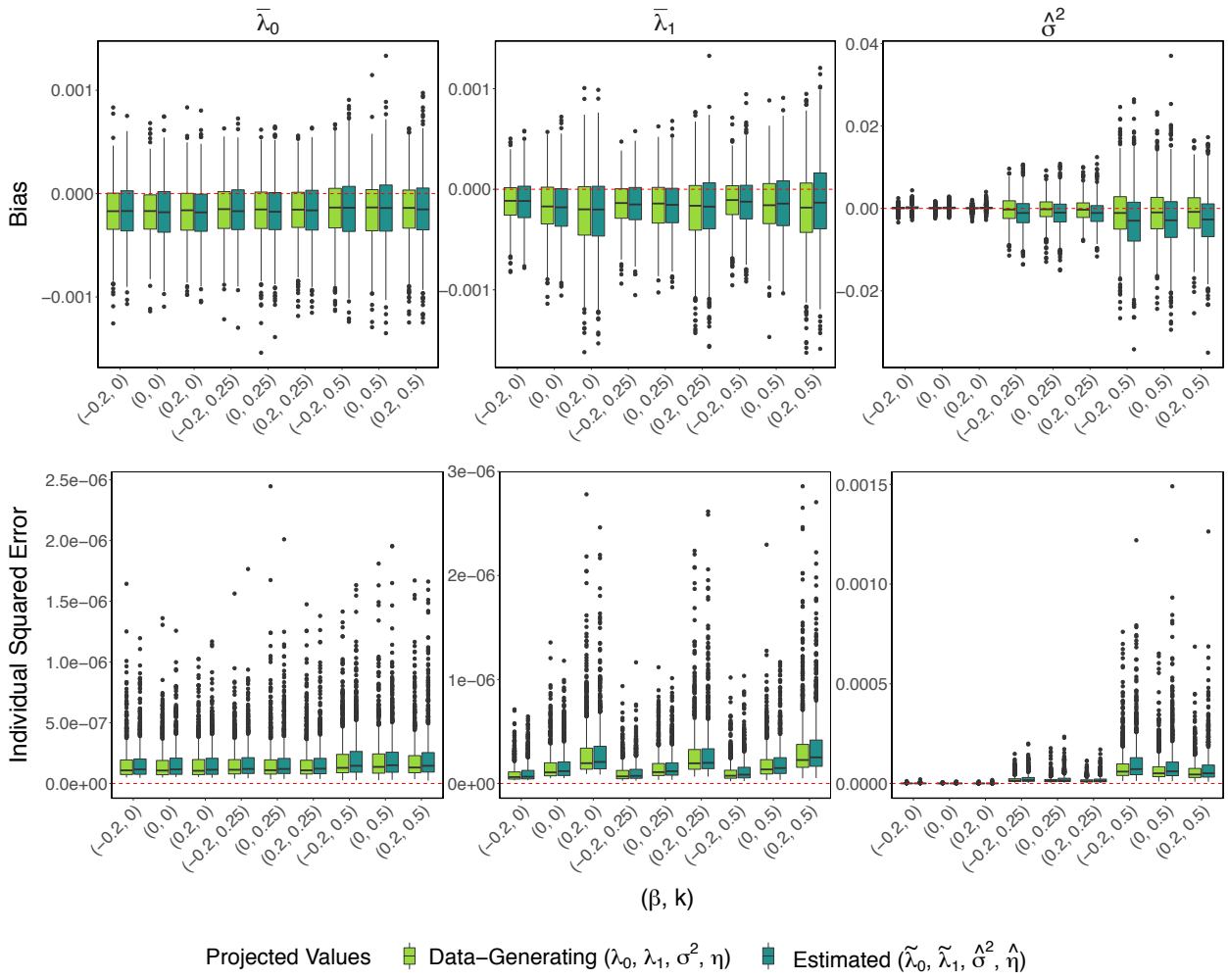

Figure S5. Bias and squared error of the estimated conditional hazards (left, center) and log-frailty variance (right) in the projected complete-trial datasets as estimators of the analogous quantities in the original completed trials. The original data were generated with $\lambda_0 = 0.01$, and selected results are presented for $\beta = (-0.2, 0.0, 0.2)$ and $k = (0.00, 0.25, 0.50)$.

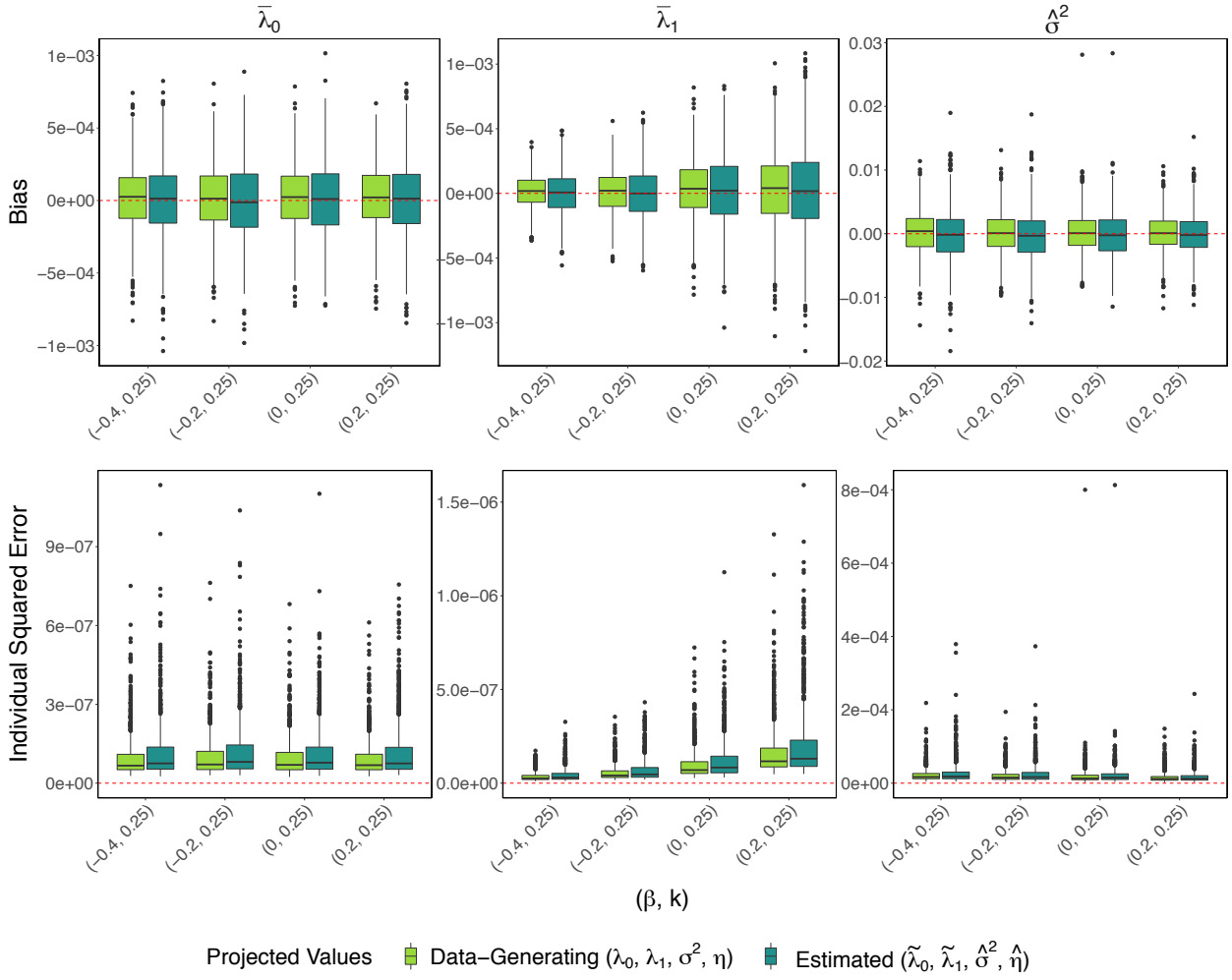

Figure S6. Sensitivity of features of the projected complete-trial datasets (considered as estimators of the analogous quantities in the original completed trials) to the width of the censoring interval under high incidence ($\lambda_0 = 0.01$) and low dependence ($k = 0.25$).

Table S1

Attributes of the cluster-conditional model, $S(t|X_i = 0)$, under lognormal frailties and either exponential or Weibull hazards. The Weibull distribution was parametrized so that $\mathbb{E}(T_i) = \lambda\Gamma(1 + 1/\kappa)$, and the parameters chosen so that the survival at 208 weeks would be the same under both distributions.

Survival Distribution		
	Exponential	Weibull
Low Incidence:	$\lambda = 0.001$	$\kappa = 0.5$, $\lambda = 4807.692$
High Incidence:	$\lambda = 0.01$	$\kappa = 2.0$, $\lambda = 144.2221$
Frailty Distribution		
Variance:	$\sigma^2 = 0.22, 0.45, 0.69$ $(k = 0.50, 0.75, 1.00)$	

Table S2

Average number of recorded events, number of person-weeks under observation, and incidence rates—stratified by arm—across 1000 original complete-trial datasets and their corresponding projected datasets. The original data were generated assuming low incidence ($\lambda_0 = 0.001$), while the projected data were generated as discussed in Sections 2.1–2.4 of the main text.

			Standard-of-Care			Intervention			e^{β}	σ^2
			Events	Person-years	$\lambda(t\|X_i)$	Events	Person-years	$\lambda(t\|X_i)$		
$k = 0.00$	$\beta = -0.2$	Original Trial	660.0	660,164.7	0.001	550.3	671,960.5	0.001	0.820	0.002
		Projected Trial	656.7	660,443.6	0.001	547.6	672,165.8	0.001	0.821	0.004
	$\beta = 0.0$	Original Trial	660.9	660,684.7	0.001	660.5	659,986.2	0.001	1.002	0.002
		Projected Trial	656.6	660,916.5	0.001	656.6	660,090.8	0.001	1.004	0.004
	$\beta = 0.2$	Original Trial	660.7	660,075.8	0.001	788.4	647,475.9	0.001	1.218	0.002
		Projected Trial	656.0	660,187.4	0.001	782.3	647,899.8	0.001	1.219	0.003
$k = 0.25$	$\beta = -0.2$	Original Trial	674.9	657,968.9	0.001	562.5	670,327.3	0.001	0.823	0.059
		Projected Trial	668.5	658,471.7	0.001	557.4	670,662.4	0.001	0.824	0.054
	$\beta = 0.0$	Original Trial	673.9	658,813.7	0.001	675.1	657,893.9	0.001	1.010	0.057
		Projected Trial	667.8	659,130.3	0.001	667.9	658,418.2	0.001	1.008	0.053
	$\beta = 0.2$	Original Trial	676.0	657,799.3	0.001	806.1	645,825.8	0.001	1.222	0.058
		Projected Trial	669.8	658,273.7	0.001	799.2	646,122.6	0.001	1.225	0.054
$k = 0.50$	$\beta = -0.2$	Original Trial	718.9	654,975.1	0.001	596.2	666,314.2	0.001	0.825	0.214
		Projected Trial	711.8	655,413.4	0.001	588.2	666,740.6	0.001	0.825	0.207
	$\beta = 0.0$	Original Trial	712.6	654,637.5	0.001	710.2	655,376.1	0.001	1.013	0.212
		Projected Trial	704.2	655,091.6	0.001	701.2	655,750.6	0.001	1.014	0.206
	$\beta = 0.2$	Original Trial	713.0	655,351.7	0.001	844.8	640,064.6	0.001	1.236	0.214
		Projected Trial	704.0	655,930.8	0.001	836.6	640,596.9	0.001	1.242	0.209

Table S3

Average number of recorded events, number of person-weeks under observation, and incidence rates—stratified by arm—across 1000 original complete-trial datasets and their corresponding projected datasets. The original data were generated assuming high incidence ($\lambda_0 = 0.01$), while the projected data were generated as discussed in Sections 2.1–2.4 of the main text.

			Standard-of-Care			Intervention			e^{β}	σ^2
			Events	Person-years	$\lambda(t\|X_i)$	Events	Person-years	$\lambda(t\|X_i)$		
$k = 0.00$	$\beta = -0.2$	Original Trial	3,242.6	331,439.5	0.010	3,012.0	373,435.4	0.008	0.819	0.000
		Projected Trial	3,231.3	332,455.1	0.010	2,998.9	374,385.9	0.008	0.820	0.001
	$\beta = 0.0$	Original Trial	3,249.2	332,008.3	0.010	3,246.9	332,280.9	0.010	0.999	0.000
		Projected Trial	3,237.3	333,014.7	0.010	3,234.6	333,296.2	0.010	0.999	0.001
	$\beta = 0.2$	Original Trial	3,244.9	332,286.3	0.010	3,460.2	293,035.6	0.012	1.223	0.000
		Projected Trial	3,232.4	333,331.4	0.010	3,450.3	293,943.1	0.012	1.223	0.000
$k = 0.25$	$\beta = -0.2$	Original Trial	3,226.4	333,493.4	0.010	3,001.9	372,955.4	0.008	0.826	0.055
		Projected Trial	3,215.5	334,485.6	0.010	2,987.9	374,018.7	0.008	0.825	0.054
	$\beta = 0.0$	Original Trial	3,228.6	332,735.5	0.010	3,218.2	332,346.2	0.010	1.002	0.055
		Projected Trial	3,216.9	333,704.8	0.010	3,207.1	333,337.0	0.010	1.003	0.054
	$\beta = 0.2$	Original Trial	3,229.1	333,238.3	0.010	3,433.0	294,450.1	0.012	1.229	0.054
		Projected Trial	3,217.9	334,190.2	0.010	3,424.3	295,334.6	0.012	1.231	0.053
$k = 0.50$	$\beta = -0.2$	Original Trial	3,168.0	334,859.3	0.010	2,945.6	372,868.1	0.008	0.825	0.201
		Projected Trial	3,159.3	335,750.2	0.010	2,933.4	373,871.9	0.008	0.824	0.198
	$\beta = 0.0$	Original Trial	3,169.2	334,672.2	0.010	3,171.1	334,249.3	0.010	1.020	0.199
		Projected Trial	3,159.9	335,574.0	0.010	3,162.4	335,068.8	0.010	1.022	0.196
	$\beta = 0.2$	Original Trial	3,180.4	334,590.4	0.010	3,370.6	297,621.0	0.013	1.249	0.195
		Projected Trial	3,171.0	335,515.0	0.010	3,365.7	298,286.7	0.012	1.257	0.192

Table S4

Sensitivity of power and conditional power results to the number and size of randomized clusters under low incidence ($\lambda_0 = 0.001$) and low dependence ($k = 0.25$). Results are summarized via the mean and empirical standard error over 1000 simulation runs.

	Power	Conditional Power		Power	Conditional Power	
		$(\widetilde{\lambda}_0,\ \widetilde{\lambda}_1,\ \widehat{\sigma}^2,\ \widehat{\boldsymbol{\eta}})$	$(\lambda_0,\ \lambda_1,\ \sigma^2,\ \boldsymbol{\eta})$		$(\widetilde{\lambda}_0,\ \widetilde{\lambda}_1,\ \widehat{\sigma}^2,\ \widehat{\boldsymbol{\eta}})$	$(\lambda_0,\ \lambda_1,\ \sigma^2,\ \boldsymbol{\eta})$
	Scenario One ($M \approx n_i$)			Scenario Two ($M \gg n_i$)		
$\beta = 0.0$	0.050	0.154 (0.269)	0.055 (0.132)	0.050	0.150 (0.247)	0.052 (0.122)
$\beta = -0.2$	0.689	0.669 (0.375)	0.692 (0.290)	0.811	0.734 (0.342)	0.796 (0.265)
$\beta = -0.4$	0.997	0.987 (0.068)	0.998 (0.012)	1.000	0.996 (0.035)	1.000 (0.002)

Table S5
Sensitivity of power and conditional power results to the width of the censoring interval under low dependence ($k = 0.25$). Results are summarized via the mean and empirical standard error over 1000 simulation runs.

	Power	Conditional Power		Power	Conditional Power	
		$(\widetilde{\lambda}_0,\ \widetilde{\lambda}_1,\ \widehat{\sigma}^2,\ \widehat{\boldsymbol{\eta}})$	$(\lambda_0,\ \lambda_1,\ \sigma^2,\ \boldsymbol{\eta})$		$(\widetilde{\lambda}_0,\ \widetilde{\lambda}_1,\ \widehat{\sigma}^2,\ \widehat{\boldsymbol{\eta}})$	$(\lambda_0,\ \lambda_1,\ \sigma^2,\ \boldsymbol{\eta})$
	$\lambda_0 = 0.001$			$\lambda_0 = 0.01$		
$\beta = 0.0$	0.050	0.112 (0.250)	0.059 (0.182)	0.050	0.053 (0.206)	0.049 (0.196)
$\beta = -0.2$	0.416	0.447 (0.417)	0.410 (0.398)	0.508	0.541 (0.460)	0.542 (0.460)
$\beta = -0.4$	0.929	0.916 (0.223)	0.936 (0.181)	0.975	0.975 (0.137)	0.976 (0.136)

Table S6

Sensitivity of power and conditional power results to misspecification of the frailty distribution under low incidence ($\lambda_0 = 0.001$) and low dependence ($\theta = 1/0.06$). Results are summarized via the mean and empirical standard error over 1000 simulation runs.

	Power	Conditional Power		
		$(\widetilde{\lambda}_0,\ \widetilde{\lambda}_1,\ \widehat{\sigma}^2,\ \widehat{\boldsymbol{\eta}})$	$(\widetilde{\lambda}_0,\ \overline{\lambda}_1,\ \widehat{\theta},\ \widetilde{\boldsymbol{\varpi}})$	$(\lambda_0,\ \lambda_1,\ \theta,\ \boldsymbol{\varpi})$
$\beta = 0.0$	0.050	0.101 (0.234)	0.104 (0.235)	0.053 (0.161)
$\beta = -0.2$	0.412	0.454 (0.422)	0.459 (0.423)	0.411 (0.402)
$\beta = -0.4$	0.927	0.894 (0.249)	0.895 (0.248)	0.917 (0.213)

Table S7
Input parameters used to generate data representative of the Botswana Combination Prevention Project.

Parameter	Value
Network Generation Parameters	
Rate of spatial mixing between communities	0.210
Variance of spatial mixing between communities	0.00651
Community size (individuals)	2,500
Proportion of community in incidence cohort	0.200
Disease Characteristics	
HIV prevalence at baseline	0.290
Prevalence of each viral load category	
Viral load: $(0, 400]$ copies/mL	0.139
Viral load: $(400, 3500]$ copies/mL	0.174
Viral load: $(3500, 10000]$ copies/mL	0.158
Viral load: $(10000, 500000]$ copies/mL	0.268
Viral load: > 50000 copies/mL	0.261
Probability of transmission per 100 person-years	
Viral load: $(0, 400]$ copies/mL	0.000
Viral load: $(400, 3500]$ copies/mL	0.045
Viral load: $(3500, 10000]$ copies/mL	0.120
Viral load: $(10000, 500000]$ copies/mL	0.140
Viral load: > 50000 copies/mL	0.230
Individual Attributes at Baseline	
Percent of males who are circumcised	0.127
Percent of individuals who use condoms regularly	0.400
Reduction in acquisition risk from circumcision	0.600
Reduction in transmission risk from regular condom use	0.800
Percent of HIV+ individuals eligible for ART at baseline (CD4 $<$ 350 cells/mm^3)	0.887
Percent of individuals on ART among those eligible	0.400
Percent of individuals with high viral load ($>$ 50000 copies/mL) that are ART naive	0.530
Intervention Components	
Linkage-to-care rates in standard-of-care communities	0.800
Linkage-to-care rates in intervention communities	0.900
Male circumcision coverage	

	Standard-of-care communities	Intervention communities
Year 1	0.314	0.464
Year 2	0.500	0.600
Year 3	0.600	0.680

HIV testing and counseling rates (among those not in the incidence cohort)

	Standard-of-care communities	Intervention communities
Year 1	0.243	0.790
Year 2	0.430	0.850
Year 3	0.470	0.900