Biostatistics (2020), **0**, 0, *pp.* 1–28
doi:10.1093/biostatistics/output

Randomization-Based Confidence Intervals for Cluster Randomized Trials

DUSTIN J. RABIDEAU*

Department of Biostatistics, Harvard University, T. H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA

RUI WANG

Department of Biostatistics, Harvard University, T. H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA and Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, 401 Park Drive, Boston, MA 02215, USA

djrabideau@mail.harvard.edu

Summary

In a cluster randomized trial (CRT), groups of people are randomly assigned to different interventions. Existing parametric and semiparametric methods for CRTs rely on distributional assumptions or a large number of clusters to maintain nominal confidence interval (CI) coverage. Randomization-based inference is an alternative approach that is distribution-free and does not require a large number of clusters to be valid. Although it is well-known that a CI can be obtained by inverting a randomization test, this requires randomization testing a non-zero null hypothesis, which is challenging with non-continuous and survival outcomes. In this paper, we propose a

*To whom correspondence should be addressed.

general method for randomization-based CIs using individual-level data from a CRT. This fast and flexible approach accommodates various outcome types, can account for design features such as matching or stratification, and employs a computationally efficient algorithm. We evaluate this method's performance through simulations and apply it to the Botswana Combination Prevention Project, a large HIV prevention trial with an interval-censored time-to-event outcome.

Key words: Cluster randomized trial; Confidence interval; Correlated data; Interval-censored; Permutation test; Randomization-based inference.

1. Introduction

In a cluster randomized trial (CRT), groups of people rather than individuals are randomly assigned to receive different interventions. The correlation between individual-level outcomes must be taken into account in the statistical analysis. With continuous, count, or binary outcomes, this is typically done using a generalized linear mixed model (GLMM) fit via likelihood-based methods (Breslow and Clayton, 1993) or a marginal model fit via a generalized estimating equation (GEE) (Liang and Zeger, 1986). Similar mixed and marginal approaches exist for correlated right-censored time-to-event outcomes (Cai and Prentice, 1995; Ripatti and Palmgren, 2000; Therneau *and others*, 2003; Wei *and others*, 1989) and interval-censored outcomes (Bellamy *and others*, 2004; Gao *and others*, 2019; Goggins and Finkelstein, 2000; Kim and Xue, 2002; Kor *and others*, 2013; Li *and others*, 2014). These parametric and semiparametric methods generally require correct specification of distributional assumptions or a large number of clusters to maintain nominal type I error and confidence interval (CI) coverage. Although various small-sample corrections have been proposed, their performance can vary depending on the particular CRT scenario (Scott *and others*, 2017; Leyrat *and others*, 2018).

Randomization-based inference is an alternative approach that is distribution-free and exact.

That is, it does not require specification of a particular parametric family of probability distributions and does not rely on a large number of clusters to be valid (Edgington, 1995; Ernst, 2004). Recently, there has been a resurgence of interest in randomization-based inference for CRTs, but this body of work has focused on hypothesis testing, not CIs.

It is well-known that, in theory, a randomization-based CI can be obtained by inverting a randomization test, i.e. by searching for and selecting the most extreme treatment effect values not rejected by the test. This approach requires carrying out randomization tests of a non-zero null hypothesis, such as $H_0 : \theta = \theta_0$ where $\theta_0 \neq 0$. With continuous outcomes and two treatment groups, this can be done by transforming all individual-level outcomes in one group by some fixed value (Edgington, 1995, Section 4.1). Beyond continuous outcomes, however, randomization testing a non-zero null—and consequently calculating a CI—is more difficult. Among the papers that have considered randomization-based CIs for non-continuous outcomes in the CRT setting, most resort to a cluster-level approach (Gail *and others*, 1996; Hughes *and others*, 2019; Raab and Butcher, 2005; Thompson *and others*, 2018). For example, Gail *and others* (1996) used randomization-based inference in a large smoking cessation CRT with a binary outcome, but used a cluster-level summary statistic to calculate a CI. This allowed them to apply the usual transformation approach to a continuous cluster-level measure. Cluster-level approaches based on unweighted summary statistics can suffer from suboptimal efficiency when cluster sizes vary substantially because they do not incorporate the differing amounts of information contributed by each cluster. Although weighting methods or incorporation of individual-level covariates can improve the efficiency of a cluster-level analysis, the former typically requires accurate estimation of the intracluster correlation coefficient, which is difficult to obtain in practice, and the latter typically requires a two-stage approach (Hayes and Moulton, 2017, Section 11.1).

Individual-level regression approaches, on the other hand, more naturally incorporate proper weights and covariate information in a single model; however, only a handful of papers have considered individual-level approaches for randomization-based CIs. For exponential family outcomes, although the duality of testing and CIs has been pointed out (Braun and Feng, 2001; Ji *and others*, 2017), no details were given as to how to carry out the necessary non-zero null hypothesis test for non-continuous outcomes. In discussing methods for covariate-adjusted randomization tests and their application to balanced CRTs, Raab and Butcher (2005) mentioned an approach for binary outcomes by introducing an offset term into a logistic model to calculate adjusted residuals, but did not pursue or develop this idea further. For time-to-event outcomes, Wang and De Gruttola (2017) briefly described in the discussion how permutation tests can be inverted to obtain interval estimates when the parameter of interest is one in an accelerated failure time model. This approach required transformation of the unknown time-to-event outcomes and cannot be used to make inference about a treatment effect based on a Cox proportional hazards model. The difficulty in constructing randomization-based CIs for non-continuous and survival outcomes can manifest in an inconsistency between the p-value and CI reported in practice, where the former is based on randomization and the latter is not. For example, in the Botswana Combination Prevention Project, a recently published large HIV prevention trial with an interval-censored outcome, the authors used a randomization test for the primary analysis, but resorted to a semiparametric Cox model to calculate a CI (Makhema *and others*, 2019).

In this paper, we propose a general method to construct randomization-based CIs for the intervention effect using individual-level data from a CRT. This fast and flexible approach accommodates various outcome types (e.g. continuous, binary, count, right- or interval-censored time-to-event), can account for design features in randomization (e.g. matching, stratification), and employs a computationally efficient algorithm. We introduce notation and present our approach in Section 2. Simulations in Section 3 demonstrate the properties of our method and compare it to alternatives. In Section 4, we re-analyze data from the Botswana Combination Prevention Project using our novel approach to obtain a randomization-based CI for the intervention effect. We conclude with a discussion in Section 5 and give a link to our R package in Section 6.

2. Methods

2.1 *Setting and notation*

Consider a CRT with two groups: intervention (or treatment) and no intervention (or control). Suppose we have a total of K clusters, n_k individuals in the kth cluster, $k = 1, \ldots, K$, and $N = \sum_{k=1}^{K} n_k$ total observations in the study. Let $X_k = 1$ indicate assignment to the intervention, $X_k = 0$ for no intervention, and $\mathbf{X} = (X_1, \ldots, X_K)^T$ denote the entire random treatment vector. Assign $K_1 = \sum_{k=1}^{K} X_k$ clusters to intervention and the remaining $K_0 = K - K_1$ to no intervention according to some predefined randomization scheme. Let $\mathbf{x} = (x_1, \ldots, x_K)^T$ denote the resulting observed treatment vector post-randomization. In this paper, we consider a wide range of outcomes including continuous, binary, and count data, as well as time-to-event data subject to right- or interval-censoring. For exponential family outcomes, let $Y_{ki},\ i = 1, \ldots, n_k$, denote the outcome random variable for the ith individual in the kth cluster. Collect all cluster-specific outcomes in a vector $\mathbf{Y}_k = (Y_{k1}, \ldots, Y_{kn_k})^T$ and all study outcomes in a vector $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_K^T)^T$.

When outcomes of interest are times-to-event, we follow the conventional notation. For right-censored data, let T_{ki} and C_{ki} denote the underlying survival time and potential censoring time for an individual measured from study entry. The observation for an individual is (U_{ki}, δ_{ki}), where $U_{ki} = \min(T_{ki}, C_{ki})$ and $\delta_{ki} = I(T_{ki} \leqslant C_{ki})$ is an indicator of whether T_{ki} is observed. We use $(\mathbf{U}_k, \boldsymbol{\delta}_k)$ and $(\mathbf{U}, \boldsymbol{\delta})$ to denote cluster-specific and all study outcomes, respectively. We assume that censoring is noninformative: that is, T and C are conditionally independent given X. For interval-censored data, we again let T_{ki} denote the time-to-event measured from study entry. We consider M distinct monitoring times $\{Z_{ki,m}, m = 1, \ldots, M\}$ at which the outcome of interest is assessed. The interval-censored outcome for an individual is $(L_{ki}, R_{ki}]$, where L_{ki} is the last observed monitoring time individual i in cluster k is event-free, and R_{ki} is the first observed monitoring time the event of interest for this individual has occurred. Individuals who remain event-free at the last observed monitoring time are right-censored with observed interval $(Z_{ki,M}, \infty)$. We use $(\mathbf{L}_k, \mathbf{R}_k]$ and $(\mathbf{L}, \mathbf{R}]$ to denote cluster-specific and all study outcomes, respectively. We again assume noninformative interval censoring conditional on X, i.e. $P(T \leqslant t | L = l, R = r, L < T \leqslant R, X = x) = P(T \leqslant t | l < T \leqslant r, X = x)$.

Throughout, our focus is on estimating the randomized intervention effect.

2.2 *Randomization-based confidence intervals*

We propose a randomization-based approach to construct a CI for the intervention effect that does not rely on parametric distributional assumptions or a large number of clusters to be valid. This method can be used to make inference about a marginal (i.e. cluster-average) or conditional (e.g. covariate-adjusted, cluster-specific) intervention effect, and applies for exponential family outcomes (Section 2.2.1) and right- or interval-censored survival outcomes (Section 2.2.2).

2.2.1 *Exponential family outcomes* Let us conceive of the sample of clusters as being representative of some hypothetical reference population of clusters. In our case, consider the population model

$$(\mathbf{Y}_k | X_k = x) \sim F(\eta_x, \boldsymbol{\phi}), \quad \eta_x = g\{E(Y_{ki} | X_k = x)\} = \mu + \theta x \,, \tag{2.1}$$

for all $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$. Here, F is an arbitrary multivariate distribution characterized by η_x and $\boldsymbol{\phi}$, a vector of nuisance parameters not affected by treatment assignment (i.e. X only affects Y through the mean model). For example, if outcomes were generated from the linear mixed model $Y_{ki} = \mu^* + \theta^* X_k + \gamma_k + \epsilon_{ki}$ with random cluster intercepts $\gamma_k \sim N(0, \sigma^2_\gamma)$ and independent residual error terms $\epsilon_{ki} \sim N(0, \sigma^2_\epsilon)$, these data would coincide with population model (2.1) where F is the multivariate normal distribution with each element of the mean vector corresponding to $\eta_x = \mu^* + \theta^* x$ and an exchangeable covariance matrix parameterized by

$\boldsymbol{\phi} = (\sigma_\gamma, \sigma_\epsilon)$. For now, we define the parameter of interest as the marginal treatment effect

$$\theta = g\{E(Y_{ki}|X_k = 1)\} - g\{E(Y_{ki}|X_k = 0)\}\,. \tag{2.2}$$

With a continuous outcome and the identity link, θ corresponds to the difference in means between populations under intervention and no intervention; with a binary outcome and the logit link, θ corresponds to the log odds ratio between these two populations; with a count outcome and the log link, θ is the log incidence rate ratio.

Invariance can be used to justify the validity of a randomization test under a population model like (2.1) (Lehmann and Romano, 2005, Section 15.2). Under population model (2.1) and $H_0 : \theta = 0$, $\eta_0 = \eta_1 = \mu$. This implies that the distribution of $(\mathbf{Y}|\mathbf{X})$ is invariant under permutations of $\mathbf{X}$ that adhere to the cluster-level randomization scheme used in the study. Thus, by Theorem 15.2.1 of Lehmann and Romano (2005), a randomization test using a statistic based on our data $\mathbf{Y}$ and $\mathbf{X}$ will be level α, the pre-specified type I error rate, and can be carried out in the following way. First, we fit the marginal model

$$g\{E(Y_{ki}|X_k^{(p)})\} = \mu + \theta X_k^{(p)}\,, \tag{2.3}$$

using the observed treatment vector $\mathbf{X}^{(1)} = \mathbf{x}$ to obtain the observed marginal treatment effect estimate, $\hat{\theta}^{(1)} = \hat{\theta}$. Then for $p = 2, \ldots, P$, we randomly permute elements of $\mathbf{x}$ in accordance with the randomization scheme, fit model (2.3) using the permuted treatment vector $\mathbf{X}^{(p)}$, and obtain a new estimate $\hat{\theta}^{(p)}$. The p-value is calculated as the proportion of $\{\hat{\theta}^{(p)}\}_{p=1}^{P}$ as or more extreme than $\hat{\theta}$. This Monte Carlo approximation to the exact p-value is used in practice because complete enumeration of all possible treatment assignment vectors is often infeasible (Dwass, 1957). We reject the null hypothesis of no treatment effect if this p-value is less than α. We can obtain $\hat{\theta}^{(p)}$ in various ways, e.g. via GEE or by fitting a generalized linear model (GLM) via maximum likelihood. An appropriate number of permutations P for the randomization test can be based on the standard error (SE) of the Monte Carlo approximation, i.e. by ensuring the SE expression

$\sqrt{a(1-a)/P}$ is adequately low for a particular p-value $= a$, or most conservatively for $a = 0.5$.

A randomization-based confidence set for θ can be obtained by inverting this randomization test. This requires testing null hypotheses of the form $H_0 : \theta = \theta_0 \in \Theta$ and collecting the set of values not rejected by these tests. For simple settings with continuous outcomes (e.g. a two-sample t-test), this can be done by subtracting the hypothesized value θ_0 from individual-level outcomes in one group, which reformulates the problem back into testing a zero null hypothesis. For linear regression settings, tests can be constructed by removing treatment effects and working with residuals. However, such approaches do not apply to non-continuous outcomes (e.g. binary, time-to-event), which are commonly used in CRTs. Our approach overcomes this challenge by working with model (2.3) directly. Rather than attempting to transform the individual-level outcomes, we carefully obtain values of the test statistic that form the randomization distribution under each non-zero null hypothesis. We can rewrite the marginal model as

$$g\{E(Y_{ki}|X_k)\} = \mu + \theta_0 X_k + (\theta - \theta_0)X_k = \mu + \theta_0 X_k + \tau X_k\,, \tag{2.4}$$

and test an equivalent null hypothesis of $H_0 : \tau = (\theta - \theta_0) = 0$. This requires obtaining estimates $\{\hat{\tau}^{(p)}\}_{p=1}^{P}$ under different randomizations, each of which can be calculated by fitting the model

$$g\{E(Y_{ki}|X_k^{(p)})\} = \mu + \theta_0 x_k + \tau X_k^{(p)}\,, \tag{2.5}$$

using the observed treatment vector $\mathbf{x}$ for the fixed offset term $\theta_0 x_k$ and the permuted treatment vector $\mathbf{X}^{(p)}$ for the offset-adjusted treatment effect term $\tau X_k^{(p)}$. Under population model (2.1) and $H_0 : \tau = 0$, the only functional relationship preventing $(\mathbf{Y}|\mathbf{X})$ from being invariant is a (transformed) mean shift of θ_0 between treatment and control clusters. By including the fixed offset term $\theta_0 x_k$ in model (2.5) across all P permutations, we eliminate this shift, resulting in invariance and, thus, a level α randomization test for any θ_0. Carrying out this test across all $\theta_0 \in \Theta$ and collecting the set of values not rejected by these tests provides a $(1-\alpha)$ randomization-based confidence set for θ, the bounds of which form a $(1-\alpha) \times 100\%$ CI for θ. This method uses individual-level data and works for a generic outcome and link function. For a continuous outcome with the identity link, this coincides exactly with the commonly used approach of transforming the outcome directly.

Note that like other randomization-based methods, because the randomization distribution of our test statistic is discrete, an exact arbitrary size α test may require a randomized testing procedure, i.e., flipping a biased coin for values on the boundary of the rejection region in order to obtain exact size α; otherwise, the test is level α and the corresponding confidence level can be conservative. For example, if there were only 50 unique equiprobable values of the test statistic in the randomization distribution, the largest achievable randomization test p-value less than $\alpha = 0.05$ is $2/50 = 0.04$; inverting this test would result in a 96% CI. This conservativeness becomes negligible as the number of values in the randomization distribution increases.

Summarizing up to this point, we have the following results. Given population model (2.1), a randomization test of $H_0 : \tau = (\theta - \theta_0) = 0$ using $\hat{\tau}$ from offset-adjusted model (2.5) is level α. Correspondingly, the set of $\theta_0 \in \Theta$ not rejected by this randomization test form a $(1 - \alpha)$ confidence set for θ, the bounds of which form a $(1 - \alpha) \times 100\%$ CI.

This CI approach is quite flexible: it can be generalized to obtain interval estimates for covariate-adjusted and cluster-specific treatment effects. Let $\mathbf{V}_{ki}$ denote a vector of cluster- or individual-level covariates. Suppose that we are interested in estimating $\theta^\dagger = g\{E(Y_{ki}|X_k = 1, \mathbf{V}_{ki})\} - g\{E(Y_{ki}|X_k = 0, \mathbf{V}_{ki})\}$, and the population model is given by

$$(\mathbf{Y}_k|X_k = x, \mathbf{V}_k) \sim F(\eta_x, \boldsymbol{\phi}), \quad \eta_x = g\{E(Y_{ki}|X_k = x, \mathbf{V}_{ki})\} = \mu^\dagger + \mathbf{V}_{ki}^T \boldsymbol{\beta}^\dagger + \theta^\dagger x\,, \qquad (2.6)$$

where $\boldsymbol{\phi}$ is not affected by treatment assignment. A level α randomization test of $H_0 : \tau^\dagger = (\theta^\dagger - \theta_0^\dagger) = 0$ can be obtained using the randomization distribution formed by $\hat{\tau}^\dagger$ from the offset-adjusted model

$$g\{E(Y_{ki}|X_k^{(p)}, \mathbf{V}_{ki})\} = \mu^\dagger + \mathbf{V}_{ki}^T \boldsymbol{\beta}^\dagger + \theta_0^\dagger x_k + \tau^\dagger X_k^{(p)}\,, \qquad (2.7)$$

because after controlling for $\mathbf{V}_{ki}$ and removing the covariate-adjusted treatment effect $\theta_0^\dagger$ via the fixed offset term, the distribution of $(\mathbf{Y}|\mathbf{X}, \mathbf{V})$ is invariant under permutations of $\mathbf{X}$. Collecting all $\theta_0^\dagger \in \Theta^\dagger$ not rejected by this test forms a $(1-\alpha)$ confidence set for $\theta^\dagger$, the bounds of which form a $(1-\alpha) \times 100\%$ CI. Similar results can be obtained for the cluster-specific treatment effect $\theta^* = g\{E(Y_{ki}|X_k = 1, \gamma_k)\} - g\{E(Y_{ki}|X_k = 0, \gamma_k)\}$ using the offset-adjusted GLMM

$$g\{E(Y_{ki}|X_k^{(p)}, \gamma_k)\} = \mu^* + \theta_0^* x_k + \tau^* X_k^{(p)} + \gamma_k \, . \tag{2.8}$$

These conditional approaches might result in improved efficiency of the randomization-based CI, but in general relate to a different (conditional, not marginal) intervention effect.

2.2.2 *Time-to-event outcomes* For right-censored data, suppose that we are interested in estimating θ, the marginal log hazard ratio of treatment in the Cox proportional hazards model $\lambda_{ki}(t|X_k = x) = \lambda_0(t)\exp(\theta x)$. Consider the population model

$$\{(\mathbf{U}_k, \boldsymbol{\delta}_k)|X_k = x\} \sim F(\lambda_x, \boldsymbol{\phi}), \quad \lambda_x = \lambda_{ki}(t|X_k = x) = \lambda_0(t)\exp(\theta x) \, , \tag{2.9}$$

where $\boldsymbol{\phi}$ is a vector of nuisance parameters not affected by treatment assignment and where censoring is noninformative. Following similar arguments to before, a level α randomization test of $H_0 : \tau = (\theta - \theta_0) = 0$ can be obtained using the randomization distribution of $\hat{\tau}$ from the offset-adjusted proportional hazards model

$$\lambda_{ki}(t|X_k^{(p)}) = \lambda_0(t)\exp(\theta_0 x_k + \tau X_k^{(p)}) \, . \tag{2.10}$$

A corresponding $(1-\alpha) \times 100\%$ CI for θ can be obtained by inverting this randomization test. Similar results apply to interval-censored data, where instead of (2.9), the population model would now become $\{(\mathbf{L}_k, \mathbf{R}_k]|X_k = x\} \sim F(\lambda_x, \boldsymbol{\phi})$, where $\lambda_x = \lambda_{ki}(t|X_k = x) = \lambda_0(t)\exp(\theta x)$.

2.2.3 *Other considerations* More generally, this randomization-based approach to CIs can be used to make inference about the intervention effect in models that accommodate an offset term.

Other than proper randomization, validity of this method relies on correct specification of the population model. This assumption is more flexible than its counterpart in mixed models because, although we similarly postulate a common distribution F across clusters, aside from the mean (or proportional hazards) model, the form of this distribution is left completely unspecified; i.e. correct specification of a particular probability distribution is not required. It can be more restrictive than other semiparametric methods (e.g., GEEs) because the population models considered here limit the effect of intervention to one aspect of the distribution (the mean for exponential family outcomes and hazard function for survival outcomes). Importantly, however, its validity does not rely on the number of clusters being large. Another advantage of randomization-based inference is the ease of accounting for design features, such as matching or stratification, in the analysis. When carrying out the randomization test, we simply consider only those treatment permutations possible under that particular design. For example, if 20 clusters were randomized in pairs, we would sample $\mathbf{X}^{(p)}$ from the $2^{10} = 1,024$ possible pair-matched randomizations as opposed to all $20!/10!10! = 184,756$ unrestricted randomizations. In these scenarios, randomization-based inference could even outperform parametric and semiparametric methods in terms of efficiency, since these alternatives often ignore such aspects of the study design.

2.3 *Computational implementation*

In addition to the difficulty of testing a non-zero null hypothesis, another stumbling block that limits the use of randomization-based CIs is the computational challenge. A standard grid or binary search requires performing randomization tests at many θ_0 to identify the bounds, each test consisting of a large number of permutations to construct the null distribution. For a typical CRT-sized data set, this could take hours or days to calculate a single CI. Instead, we adapted an efficient search procedure (Garthwaite, 1996) to our offset-adjusted approach, which can reduce the computation time down to seconds or minutes. At each step of this sequential search for the lower or upper bound, the estimate is updated based on only a single permutation of the data—thus, a single model fit. As the number of steps increases, estimates converge in probability to the correct CI bounds, i.e. the bounds that would be obtained if we carried out a full randomization test at every possible $\theta_0 \in \Theta$ (Garthwaite, 1996).

More specifically, suppose we carry out a P-step search for U, the correct upper confidence limit of θ defined in (2.2) (note, this P could be different from that we used for the randomization test). At the pth step of the search, we fit model (2.5) with the current value of the upper limit $\theta_0 = U^{(p)}$ and permuted treatment vector $\mathbf{X}^{(p)}$ to obtain the permuted offset-adjusted estimate $\hat{\tau}^{(p)}(\mathbf{X}^{(p)})$. We also directly calculate the observed offset-adjusted estimate $\hat{\tau}^{(p)}(\mathbf{x}) = \hat{\theta} - U^{(p)}$ based on the initial fit of model (2.3). We update the upper limit based on whether the permuted estimate is larger than the observed estimate

$$U^{(p+1)} = \begin{cases} U^{(p)} - \frac{c(\alpha/2)}{p}, & \text{if } \hat{\tau}^{(p)}(\mathbf{X}^{(p)}) > \hat{\tau}^{(p)}(\mathbf{x}) \\ U^{(p)} + \frac{c(1-\alpha/2)}{p}, & \text{if } \hat{\tau}^{(p)}(\mathbf{X}^{(p)}) \leqslant \hat{\tau}^{(p)}(\mathbf{x})\,, \end{cases} \tag{2.11}$$

where $c > 0$ is a chosen step length constant. On average, this corresponds to stepping down if the randomization-based p-value of $H_0 : \tau = 0$ versus $H_1 : \tau < 0$ is smaller than $\alpha/2$ and stepping up if it is greater than $\alpha/2$. This makes sense, since U would correspond to a p-value of exactly $\alpha/2$ for this one-sided randomization test. An independent search is carried out for the correct lower limit L in a similar fashion using

$$L^{(p+1)} = \begin{cases} L^{(p)} + \frac{c(\alpha/2)}{p}, & \text{if } \hat{\tau}^{(p)}(\mathbf{X}^{(p)}) < \hat{\tau}^{(p)}(\mathbf{x}) \\ L^{(p)} - \frac{c(1-\alpha/2)}{p}, & \text{if } \hat{\tau}^{(p)}(\mathbf{X}^{(p)}) \geqslant \hat{\tau}^{(p)}(\mathbf{x})\,. \end{cases} \tag{2.12}$$

To avoid early steps changing dramatically in size, the P-step search should begin with $p = m$ with $m = \min\{\lceil 0.3(4-\alpha)/\alpha \rceil, 50\}$ (Garthwaite, 1996). Thus, $[L^{(m)}, U^{(m)}]$ correspond to our chosen starting values and the final updated values $[L^{(m+P)}, U^{(m+P)}]$ are adopted as the CI.

Though this efficient search algorithm substantially reduces the computational burden of randomization-based CIs, its performance could be affected by the starting values, step length, or total number of steps chosen. Detailed guidance on choosing these tuning parameters can be found in Section 3 of Garthwaite (1996). Based on their recommendations, we use $c = k(U^{(p)} - \hat{\theta})$ and $c = k(\hat{\theta} - L^{(p)})$ for the upper and lower bound search, respectively, where $k = 2/\{z_{\alpha/2}(2\pi)^{-1/2}\exp(-z^2_{\alpha/2}/2)\}$ and $z_{\alpha/2}$ is the upper $100(\alpha/2)\%$ point of the standard normal distribution. Starting values are based on $\hat{\theta} \pm \{(t_1 + t_2)/2\}$, where t_1 and t_2 are the second smallest and second largest permuted offset-adjusted estimates from a randomization test of $H_0 : \theta = \hat{\theta}$ using $\lceil (4 - \alpha)/\alpha \rceil$ permutations. Diagnostic plots (e.g. Figure 3) can be used to guide (or confirm) adequate choice of P. We can also choose P based on the asymptotic variance of the coverage of a P-step search

$$V_p \approx \alpha(1 - \alpha/2)/P\varepsilon\,, \tag{2.13}$$

where $\varepsilon \approx 0.75$ for the recommended value of k above. Using (2.13), we can either choose an acceptable V_P and solve for P, or choose P and ensure this results in an acceptable V_P. Finally, we echo some practical suggestions made by Garthwaite (1996), such as monitoring convergence of the CI bounds, restarting the procedure if the starting values seem poor, using multiple chains with different starting values, and considering different step length constants—all of which can be implemented using our R package (see Section 6).

For longer searches (e.g. $P \geqslant 200,000$), Garthwaite and Jones (2009) proposed an improvement on this algorithm by taking larger steps during later phases of the search and averaging, rather than using only the final values, for CI estimation. Details on this extension can be found in the supplementary material available at *Biostatistics* online. In addition to the procedure outlined above, we implemented this alternative search for our data example in Section 4.

3. Simulations

We carried out simulations to evaluate the performance of this randomization-based approach and compare it to alternative methods. Simulations were run in R 3.4.1 or higher. Results for each simulation scenario were based on 10,000 independently generated data sets.

3.1 *Binary outcome*

First, we considered a parallel CRT with a binary outcome. Data were generated from the GLMM

$$\text{logit}\{E(Y_{ki}|X_k, \gamma_k)\} = \mu^* + \theta^* X_k + \gamma_k \,, \tag{3.14}$$

where μ^* was the cluster-conditional log odds of the outcome in the control group, and θ^* was the cluster-conditional log odds ratio associated with treatment. We drew random cluster effects from a normal distribution centered at zero, i.e. $\gamma_k \sim N(0, \sigma^2)$. We set the population-level prevalence of the outcome in the control group at 25% ($\mu^* = \text{logit}(0.25)$). We examined different intracluster correlation coefficients ($\sigma = 0.1,\ 0.2,\ 0.5$, which induced ICC $\approx 0.001,\ 0.01,\ 0.05$), underlying treatment effects ($\theta^* = 0,\ 0.25,\ 0.5$), numbers of clusters ($K = 10, 20, 30, 50$ with $K_0 = K_1 = K/2$), and ranges of cluster sizes (n_k drawn uniformly from 10 to 50 and 100 to 200). Each true marginal treatment effect θ induced by data generating model (3.14) was approximated using Gauss-Hermite quadrature.

We analyzed each data set with our proposed randomization-based method. We targeted the marginal treatment effect θ defined in (2.2) by using $\hat{\theta}$ from model (2.3) for the randomization test of no treatment effect and $\hat{\tau}$ from model (2.5) for the randomization-based CI, both with $g = \text{logit}$. The p-value and each bound of the 95% CI were based on $P = 5,000$, which provided sufficiently accurate p-value estimates (e.g. SE ≈ 0.003 for p-value $= 0.05$) and acceptable coverage precision of the CI search based on (2.13) (e.g. $V_p \approx 1.3 \times 10^{-5}$ for 95% coverage). We also considered three alternative approaches. First, we fit a marginal model via GEE using the standard sandwich variance estimator. We also fit a small-sample adjusted GEE using the δ_5 adjustment proposed by Fay and Graubard (2001), which has been shown to perform well in CRTs with a small number of clusters (Scott *and others*, 2017). Both GEEs targeted the marginal treatment effect θ as well. Finally, we fit GLMM (3.14) via maximum likelihood, which targeted the cluster-conditional treatment effect θ^* instead. We used a Laplace approximation (default in `lme4::glmer()` in R) to approximate the log likelihood. All three of these alternative models were correctly specified and used two-sided Wald tests and corresponding 95% CIs. CI coverages were calculated with respect to the true value of the target parameter, i.e. θ for randomization-based inference and GEE, and θ^* for GLMM.

Results for ICC ≈ 0.01 are presented in Figure 1. In general, randomization-based inference resulted in nominal type I error rates (0.046 to 0.050), nominal CI coverages (0.946 to 0.954), and moderate CI widths (0.31 to 1.45). The small-sample adjusted GEE resulted in conservative type I error rates (0.025 to 0.048), nominal to conservative CI coverages (0.948 to 0.978), and wider CIs (0.31 to 2.19), especially when $K = 10$ and cluster sizes were small. The GLMM and standard GEE both had inflated type I error (e.g. 0.10 and 0.12, respectively, for $K = 10$ and n_k from 100 to 200), undercoverage (e.g. 0.90 and 0.88), and the narrowest CIs.

Results for ICC ≈ 0.001 and 0.05 are presented in Figures S1-S2 of the supplementary material available at *Biostatistics* online. Both GEE approaches got worse as ICC and cluster sizes decreased: the standard GEE approach became more liberal while the small-sample adjusted GEE became more conservative. GLMM size and coverage became more liberal as ICC and cluster sizes increased. Randomization-based inference performed well for all settings we examined.

In terms of efficiency, our randomization-based approach outperformed the small-sample adjusted GEE, especially with smaller numbers of clusters, cluster sizes, and ICC. Across all scenarios, it provided equivalent or better power and CI widths while maintaining nominal type I error and coverage. While the GLMM and unadjusted GEE methods appeared to yield better efficiency, this result was distorted by their inflated type I error rates and undercoverage. In fact, in scenarios where the asymptotic approaches had close to nominal type I error and coverage (e.g. $K = 50$ or ICC ≈ 0.001 for GLMM), randomization-based inference resulted in very minimal efficiency loss compared to these alternative methods.

Computation times in R varied by simulation scenario, but ranged anywhere from about 20 seconds ($N \approx 300$) to 4 minutes ($N \approx 7,500$) to calculate a single randomization-based CI with $P = 5,000$ permutations, run in parallel across 2 logical CPUs on a MacBookAir8,1 with a 1.6 GHz Intel Core i5 (see Table S1 in the supplementary material available at *Biostatistics* online).

3.2 *Time-to-event outcome*

Next, we considered a pair-matched CRT with an interval-censored time-to-event outcome. This simulation was designed to align closely with our data example in Section 4, the Botswana Combination Prevention Project. Event times were generated from an exponential frailty model

$$\lambda_{jki}(t|X_{jk}, \gamma_{jk}, \eta_j) = \lambda_0(t)\exp(\theta^* X_{jk} + \gamma_{jk} + \eta_j)\,, \tag{3.15}$$

where now $j = 1, \ldots, K/2$ and $k \in \{1, 2\}$ jointly index each cluster, γ_{jk} is a cluster-specific random effect, and η_j is a pair-specific random effect, both drawn independently from a $N(0, \sigma^2)$. Actual study visit times were drawn uniformly within a two-month window centered around an annually scheduled visit, resulting in an interval-censored outcome. Loss to follow-up occurred at a constant exponential rate of about 10% each year and individuals were followed for at most three years. Each data set had $K = 30$ clusters with sizes ranging from 250 to 350 individuals.

We examined coverage and efficiency of our randomization-based approach across different underlying treatment effects ($\theta^* = 0, -0.2, -0.4, -0.8$) and within-cluster and -pair correlations ($\sigma = 0.2, 0.5, 0.8$). Randomization-based inference was based on $\hat{\theta}$, the estimated log hazard ratio from an interval-censored Weibull regression model

$$\lambda_{jki}(t|X_{jk}^{(p)}) = \lambda_0(t)\exp(\theta X_{jk}^{(p)})\,. \tag{3.16}$$

We sampled $P = 5,000$ permutations from all possible $2^{15} = 32,768$ pair-matched randomizations. For comparison, we fit two Weibull frailty models: the first corresponded to model (3.15) without η_j (i.e. only accounting for within-cluster correlation), the second without γ_{jk} (i.e. only accounting for within-pair correlation). This was done because there is currently no reliable software in R to fit an interval-censored survival model with more than one frailty term for data sets with large cluster sizes. Wald tests and CIs were used for these alternative approaches. We note that the treatment effect parameter θ in marginal model (3.16) or in either single-frailty conditional model is generally different from θ^* in the data generating model (3.15). The true values for these parameters were obtained via simulation. We also point out that data generation model (3.15), a proportional hazards model conditional on both frailties, does not necessarily induce a proportional hazards model marginally or conditional on only a single frailty (Martinussen and Andersen, 2018; Ritz and Spiegelman, 2004). In other words, all three analysis models are misspecified, making the precise interpretation of the target parameter difficult. As such, the results presented here illustrate the robustness of each method to model misspecification.

Results for $\sigma = 0.5$ are presented in Figure 2. Across all scenarios, randomization-based inference resulted in nominal type I error (0.048 to 0.053) and coverage (0.935 to 0.952). The Weibull model with a cluster-specific frailty resulted in conservative type I error (0.012 to 0.037) and coverage (0.963 to 0.988), while the model with a pair-specific frailty resulted in drastically inflated type I error (0.082 to 0.592) and severe undercoverage (0.408 to 0.957). Type I error, coverage, and convergence of both frailty models generally got worse as the within-cluster and -pair correlations increased (see Figures S3-S4 and Table S2 in the supplementary material available at *Biostatistics* online). Of note, even these relatively simple parametric survival models with a single frailty term had trouble converging up to about 15% of the time for these moderately-sized CRT data sets, whereas randomization-based CIs could always be calculated. For this larger ($N \approx 9,000$) and more complex CRT setting, it took approximately 8 minutes to compute a single randomization-based CI in R with $P = 5,000$ permutations (computed in parallel across 2 logical CPUs on a MacBookAir8,1 with a 1.6 GHz Intel Core i5).

4. Application to the Botswana Combination Prevention Project

The Botswana Combination Prevention Project (BCPP) was a pair-matched CRT to test whether a combination prevention intervention package could reduce population-level cumulative HIV incidence over three years of follow-up (Makhema *and others*, 2019). A total of 30 communities were randomized: 15 to the intervention arm (combination prevention package) and 15 to the control arm (enhanced standard of care). Randomization was done using a pair-matched design with community matching based on population size and age structure, existing health services, and geographic location. The primary individual-level outcome was time to HIV infection, measured at annual study visits within a cohort of individuals identified as HIV-negative among a 20% random sample of eligible households at baseline. That is, we had an interval-censored time-to-event outcome for each cohort participant. The HIV-incidence cohort was composed of 8,551 participants with a median of 308 (minimum 106, maximum 392) from each of the 30 communities. There were a total of 147 HIV infections, 57 in the intervention group (annualized HIV incidence, 0.59%) and 90 in the control group (annualized HIV incidence, 0.92%). The median follow-up time was 29 months. The prespecified unadjusted primary analysis, which was based on a randomization test, yielded a p-value of 0.09. The estimated hazard ratio corresponding to treatment based on a pair-stratified Cox model was 0.65 with a 95% CI of [0.46, 0.90].

We targeted the marginal log hazard ratio θ from model (3.16) and applied our randomization-based approach to the BCPP data. To confirm an adequate choice of P, in addition to considering p-value accuracy and coverage precision, we monitored four separate chains of length $5,000$, each using different initial values: first using our chosen values, second using initial values based on an asymptotic Wald CI ignoring within-cluster correlation, and third (and fourth) using initial values wider (and narrower) than the final values from the first chain. The four chains for each bound converged towards each other, began to gently oscillate around similar values, and all ended within 0.025 of each other on the hazard ratio scale (see Figure 3). Based on these characteristics, we deemed that $P = 5,000$ resulted in acceptable convergence and demonstrated adequate robustness to different starting values for these data. It was computationally feasible, however, for us to increase to $P = 20,000$ to allow for even further refinement of the final CI limits (see Figure S5 in the supplementary material available at *Biostatistics* online). This resulted in an estimated hazard ratio of 0.64 (95% CI [0.37, 1.04], p-value = 0.06). Calculation of the CI based on 20,000 permutations for each bound took about 40 minutes to run in R (computed in parallel across 2 logical CPUs on a MacBookAir8,1 with a 1.6 GHz Intel Core i5). Similar to our simulations in Section 3.2, we also fit two separate Weibull frailty models to account for the within-community and -pair correlation. Respectively, these approaches resulted in estimated hazard ratios of 0.64 (95% CI [0.43, 0.95], p-value = 0.03) and 0.64 (95% CI [0.45,0.90], p-value = 0.01). Finally, to align closely with the CI approach taken in the BCPP, we considered a pair-stratified Weibull model to estimate the treatment effect. Randomization-based inference for this model yielded a 95% CI of [0.37, 1.05] (p-value = 0.07), while the likelihood-based analysis produced [0.46, 0.91] (p-value = 0.01), both with point estimates of 0.65. All of these results are summarized in Table 1. Comparable CIs were found using the improved search procedure in Garthwaite and Jones (2009) (see Figure S6 and Table S3 in the supplementary material available at *Biostatistics* online).

Consistent with the results in Makhema *and others* (2019), the randomization-based p-values and CIs did not reach the pre-specified 0.05 level, while those based on models with stronger parametric assumptions or asymptotic approximations did. This discrepancy could be due to the usual robust-efficiency trade-off between distribution-free and parametric likelihood-based methods. As a threshold of 0.05 is not a magic number, all of these methods do provide evidence supporting an effect of the intervention on reducing HIV incidence. Nevertheless, it would have been desirable to report a p-value and CI using the same randomization-based analysis method. This was not possible at the time of the BCPP analysis, however, due to the unavailability of methods to obtain a randomization-based CI for correlated interval-censored survival outcomes.

5. Discussion

In this paper, we proposed a fast and flexible approach to randomization-based CIs using individual-level data from a CRT. We demonstrated that this method has good properties and performs well compared to other methods, even when the modeling assumptions of those alternatives were met. Randomization-based inference is especially attractive for analyzing CRTs, as it does not require specification of a particular distribution and does not rely on a large number of clusters to maintain nominal type I error and CI coverage. Another advantage of randomization-based inference is the ease with which one can account for restricted designs simply by limiting the set of treatment permutations considered in the analysis.

Some have raised concern with randomization-based inference when a different number of clusters are randomized to each arm (Braun and Feng, 2001; Gail *and others*, 1996). It is important to clarify that a randomization test does guarantee nominal type I error—even with a different number of clusters in each arm—under a null hypothesis corresponding to entire distributions being identical, e.g. $H_0 : \theta = 0$ under population model (2.1). This may not be the case, however, when testing a weaker null hypothesis corresponding to only some components of the distributions being equal. For example, if treatment does affect the cluster variances (e.g. where some component of $\boldsymbol{\phi}$ in population model (2.1) depends on treatment), then a randomization test of the null hypothesis of equal means and its corresponding CI would be liberal when the arm with fewer clusters has larger variance and conservative when it has smaller variance (Gail *and others*, 1996; Romano, 1990). This is an important distinction to keep in mind, especially if imbalanced treatment allocations and differences in variances are expected to be substantial.

As the number of clusters or size of the clusters becomes large, randomization-based CIs using individual-level data could become computationally burdensome. With a large number of small clusters (e.g. households), this is luckily where alternative semiparametric approaches like GEE perform reasonably well. With a small number of large clusters (e.g. entire communities), most of the computational burden may be alleviated by using a cluster-level randomization-based analysis, as many have suggested (Gail *and others*, 1996; Hughes *and others*, 2019; Raab and Butcher, 2005; Thompson *and others*, 2018). An unweighted cluster-level analysis may lead to some loss of statistical efficiency if cluster sizes vary substantially, although this impact would be minimal when all clusters have large sizes.

More than two (say, T) treatment groups could be accommodated in a randomization test, for example, by replacing $\theta X_k^{(p)}$ in model (2.3) with $(\mathbf{X}_k^{(p)})^T\boldsymbol{\theta}$, where $\mathbf{X}_k^{(p)}$ and $\boldsymbol{\theta}$ are vectors of length $T-1$. Similarly for the CI, we would fit model (2.5) replacing $\theta_0 x_k$ with $(\mathbf{x}_k)^T\boldsymbol{\theta}_0$ for the fixed offset term and $\tau X_k^{(p)}$ with $(\mathbf{X}_k^{(p)})^T\boldsymbol{\tau}$ for the permuted offset-adjusted treatment effect term. In theory this seems to be a straightforward extension, but in practice this means the CI search procedure must take place over a $T-1$ dimensional space, which introduces further complexity in computing. Complex CRT designs (e.g. stepped wedge) and other outcome types and regression models could be handled by modifying the model, the permutation procedure, or both.

Finally, we have focused our work on CRTs, but this randomization-based CI approach using offset adjustment can be used with independent outcomes or in other correlated data settings, such as randomized clinical trials with repeated measurements. For survival outcomes, we focused on a proportional hazards model, but the method can be applied to an accelerated failure time model provided the corresponding population model is correctly specified.

6. Software

An R package for our method is available online at `https://github.com/djrabideau/permuter`.

Supplementary Material

Supplementary material is available online at `http://biostatistics.oxfordjournals.org`.

Acknowledgments

We thank the Editors, Associate Editor, and reviewers for their helpful comments, which improved the paper. We thank the participants and study team of the Botswana Combination Prevention Project. We are also grateful to Michael D. Hughes, Judith J. Lok, Kaitlyn Cook, and Lee Kennedy-Shaffer for helpful discussions. This work was supported by T32 AI007358 from the National Institutes of Health and R01 AI136947 from the National Institute of Allergy and Infectious Diseases. *Conflict of Interest*: None declared.

References

Bellamy, S. L., Li, Y., Ryan, L. M., Lipsitz, S., Canner, M. J. and Wright, R. (2004). Analysis of clustered and interval censored data from a community-based study in asthma. *Statistics in Medicine* **23**, 3607–3621.

Braun, T. M. and Feng, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association* **96**, 1424–1432.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**, 151–164.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics* **28**, 181–187.

Edgington, E. S. (1995). *Randomization tests*, 3rd edition. New York: Dekker.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science* **19**, 676–685.

Fay, M. P. and Graubard, B. I. (2001). Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics* **57**, 1198–1206.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B. and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* **15**, 1069–1092.

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

Garthwaite, P. H. (1996). Confidence intervals from randomization tests. *Biometrics* **52**, 1387–1393.

Garthwaite, P. H. and Jones, M. C. (2009). A stochastic approximation method and its application to confidence intervals. *Journal of Computational and Graphical Statistics* **18**, 184–200.

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

Hayes, R. J. and Moulton, L. H. (2017). *Cluster Randomised Trials*, 2nd edition. New York: Chapman and Hall/CRC.

Hughes, J. P., Heagerty, P. J., Xia, F. and Ren, Y. (2019). Robust inference for the stepped wedge design. *Biometrics* (in press).

Ji, X., Fink, G., Robyn, P. J. and Small, D. S. (2017). Randomization inference for stepped-wedge cluster-randomized trials: an application to community-based health insurance. *The Annals of Applied Statistics* **11**, 1–20.

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

Kor, C., Cheng, K. and Chen, Y. (2013). A method for analyzing clustered interval-censored data based on cox's model. *Statistics in Medicine* **32**, 822–832.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, 3rd edition. New York: Springer.

Leyrat, C., Morgan, K. E., Leurent, B. and Kahan, B. C. (2018). Cluster randomized trials with a small number of clusters: which analyses should be used? *International Journal of Epidemiology* **47**, 321–331.

Li, J., Tong, X. and Sun, J. (2014). Sieve estimation for the cox model with clustered interval-censored failure time data. *Statistics in Biosciences* **6**, 55–72.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Makhema, J., Wirth, K. E., Pretorius Holme, M., Gaolathe, T., Mmalane, M., Kadima, E., Chakalisa, U., Bennett, K., Leidner, J., Manyake, K., Mbikiwa, A. M., Simon, S. V., Letlhogile, R., Mukokomani, K., van Widenfelt, E., Moyo, S., Lebelonyane, R., Alwano, M. G., Powis, K. M., Dryden-Peterson, S. L., Kgathi, C., Novitsky, V., Moore, J., Bachanas, P., Abrams, W., Block, L., El-Halabi, S., Marukutira, T., Mills, L. A., Sexton, C., Raizes, E., Gaseitsiwe, S., Bussmann, H., Okui, L., John, O., Shapiro, R. L., Pals, S., Michael, H., Roland, M., De Gruttola, V., Lei, Q., Wang, R., Tchetgen Tchetgen, E., Essex, M. *and others*. (2019). Universal testing, expanded treatment, and incidence of hiv infection in botswana. *New England Journal of Medicine* **381**, 230–242.

Martinussen, T. and Andersen, S. Vansteelandt P. K. (2018). Subtleties in the interpretation of hazard ratios. *ArXiv*.

Raab, G. M. and Butcher, I. (2005). Randomization inference for balanced cluster-randomized trials. *Clinical Trials* **2**, 130–140.

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

Ritz, J. and Spiegelman, D. (2004). Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research* **13**, 309–323.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association* **85**, 686–692.

Scott, J. M., deCamp, A., Juraska, M., Fay, M. P. and Gilbert, P. B. (2017). Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. *Statistical Methods in Medical Research* **26**, 583–597.

Therneau, T., Grambsch, P. and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics* **12**, 156–175.

Thompson, J. A., Davey, C., Fielding, K., Hargreaves, J. R. and Hayes, R. J. (2018). Robust analysis of stepped wedge trials using cluster-level summaries within periods. *Statistics in Medicine* **37**, 2487–2500.

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.

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

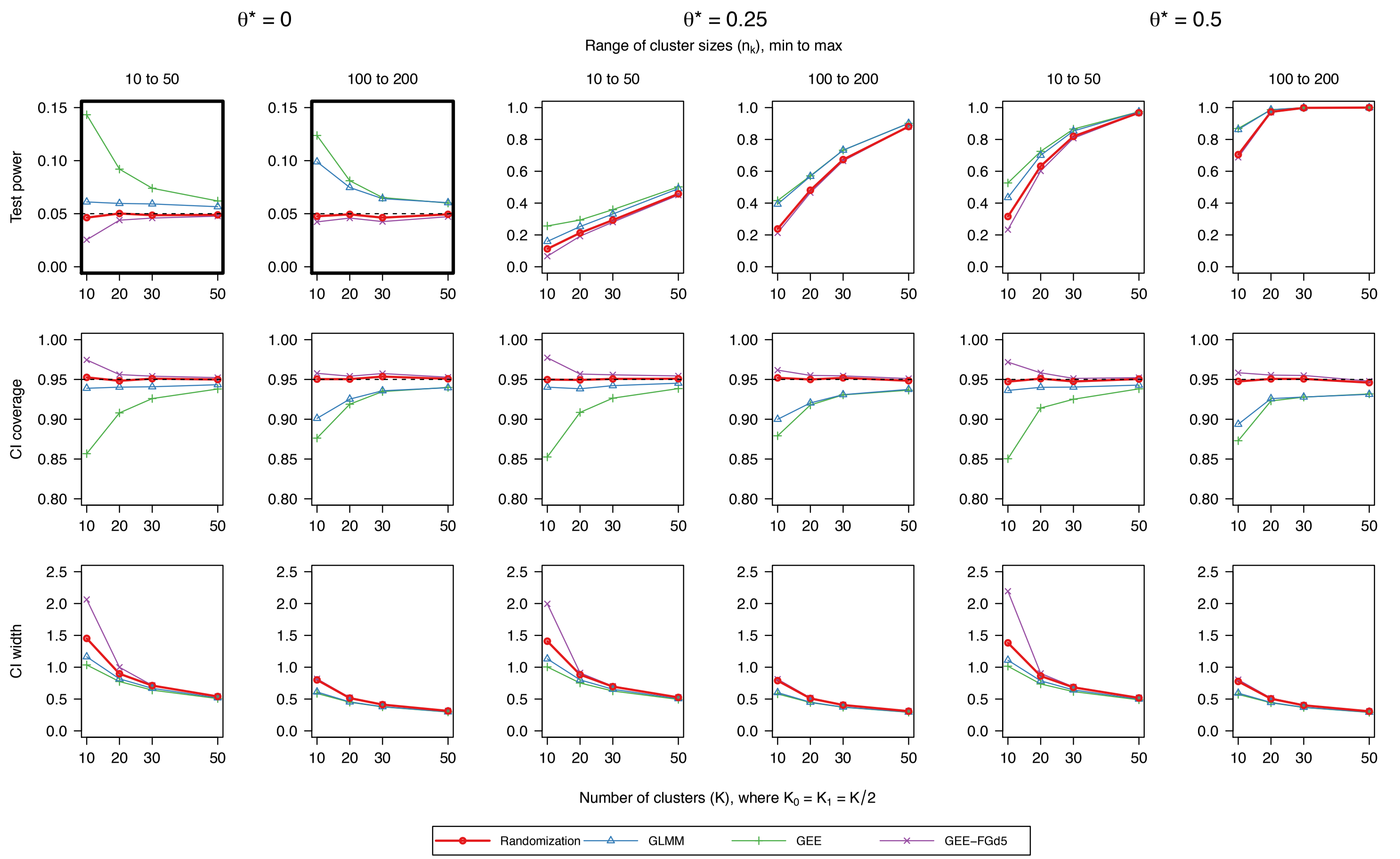

Fig. 1. Simulation results of a parallel CRT with a binary outcome (ICC $\approx$ 0.01). Methods considered include randomization-based inference (Randomization), a logistic generalized linear mixed model fit via maximum likelihood (GLMM), a marginal model fit via a generalized estimating equation (GEE), and a small-sample adjusted GEE (GEE-FGd5). Two bold upper-left plots have a different y-axis range from others in the first row for clarity. This figure appears in color in the electronic version of this article.

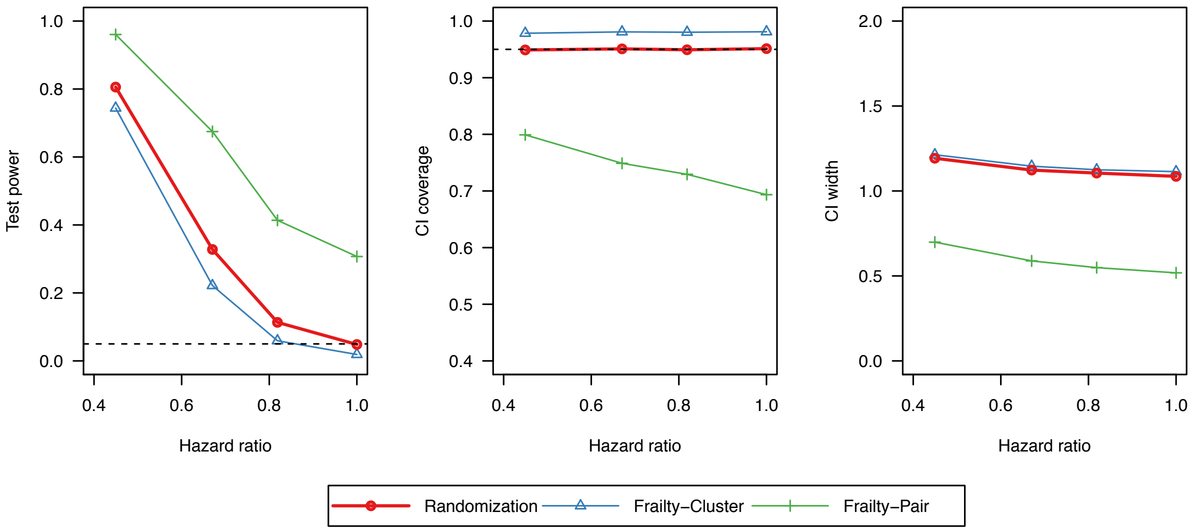

Fig. 2. Simulation results of a pair-matched CRT with an interval-censored time-to-event outcome ($\sigma = 0.5$). Methods considered included randomization-based inference (Randomization), a Weibull model with a frailty term for community (Frailty-Cluster), and with a frailty term for pair (Frailty-Pair). This figure appears in color in the electronic version of this article.

Table 1. *Estimated hazard ratio (HR) of the intervention effect in the Botswana Combination Prevention Project. Methods considered include randomization-based inference (Randomization), both marginal and pair-stratified, a Weibull model with a frailty term for community (Weibull, Frailty-Cluster), with a frailty term for pair (Weibull, Frailty-Pair), and a pair-stratified Weibull model (Weibull, Pair-Stratified).*

Method	HR	95% CI	p-value
Randomization, Marginal	0.640	[0.374, 1.039]	0.064
Randomization, Pair-Stratified	0.646	[0.369, 1.054]	0.068
Weibull, Frailty-Cluster	0.640	[0.432, 0.947]	0.025
Weibull, Frailty-Pair	0.641	[0.453, 0.905]	0.012
Weibull, Pair-Stratified	0.646	[0.457, 0.913]	0.013

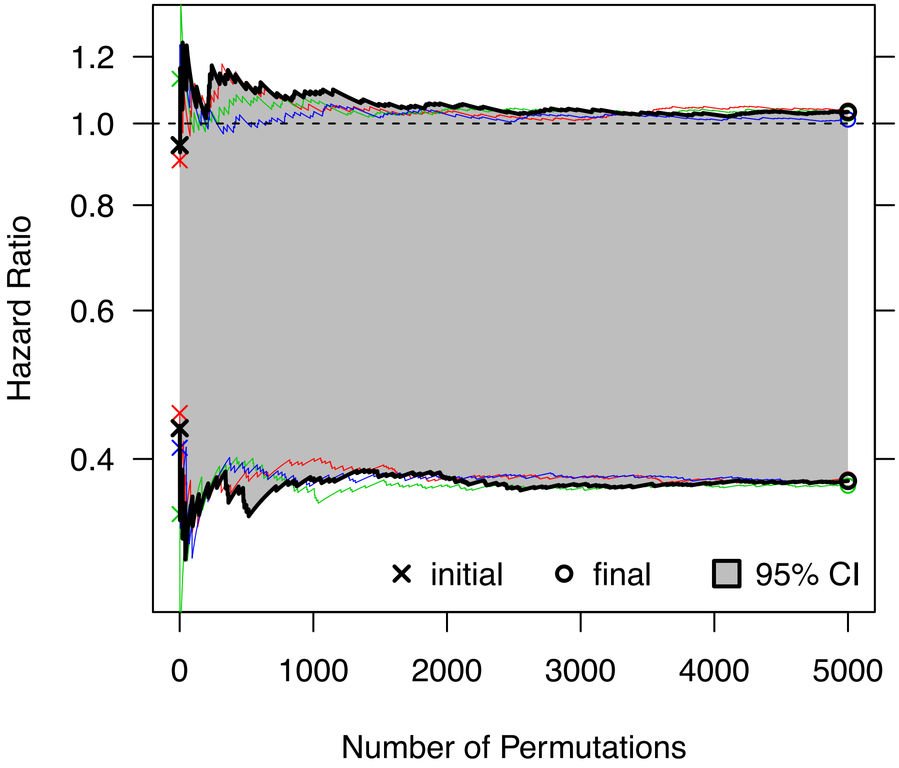

Fig. 3. Efficient search for a randomization-based 95% confidence interval (CI) for the intervention effect in the Botswana Combination Prevention Project, using the procedure adapted from Garthwaite (1996). Here we monitor four separate chains using different starting values. The first 5,000 (of 20,000 total) steps of our final CI estimate are indicated by bold black lines; other chains (included to assess adequate convergence) are indicated by thin colored lines. This figure appears in color in the electronic version of this article.

Supplementary Material to Randomization-Based Confidence Intervals for Cluster Randomized Trials

DUSTIN J. RABIDEAU*

Department of Biostatistics, Harvard University, T. H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA

RUI WANG

Department of Biostatistics, Harvard University, T. H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA and Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, 401 Park Drive, Boston, MA 02215, USA

djrabideau@mail.harvard.edu

A. Alternative Procedure for Longer CI Searches, Adapted from Garthwaite and Jones (2009)

For longer searches (e.g. $P \geqslant 200,000$), Garthwaite and Jones (2009) proposed an improvement on the algorithm in Garthwaite (1996) by taking larger steps during later phases of the search and averaging, rather than using only the final values, for CI estimation. This alternative procedure also adapts to our offset-adjusted approach.

Specifically, suppose we carry out a P-step search for each bound, broken into three separate phases, with $P = P_1 + P_2 + P_3$. Starting values $[L^{(m)}, U^{(m)}]$ for Phase 1 and step length constant c throughout can be chosen as outlined in Section 2.3. Updates are made using the same general formulas

$$U^{(p+1)} = \begin{cases} U^{(p)} - a_p(\alpha/2), & \text{if } \widehat{\tau}^{(p)}(\mathbf{X}^{(p)}) > \widehat{\tau}^{(p)}(\mathbf{x}) \\ U^{(p)} + a_p(1-\alpha/2), & \text{if } \widehat{\tau}^{(p)}(\mathbf{X}^{(p)}) \leqslant \widehat{\tau}^{(p)}(\mathbf{x}) \end{cases}$$

$$L^{(p+1)} = \begin{cases} L^{(p)} + a_p(\alpha/2), & \text{if } \widehat{\tau}^{(p)}(\mathbf{X}^{(p)}) < \widehat{\tau}^{(p)}(\mathbf{x}) \\ L^{(p)} - a_p(1-\alpha/2), & \text{if } \widehat{\tau}^{(p)}(\mathbf{X}^{(p)}) \geqslant \widehat{\tau}^{(p)}(\mathbf{x})\,, \end{cases}$$

but using a different step length a_p in each phase:

$$a_p = \begin{cases} c/p, & \text{(Phase 1)} \quad p = m, \ldots, m + P_1 \\ c/(m + P_1), & \text{(Phase 2)} \quad p = m + P_1, \ldots, m + P_1 + P_2 \\ c/\{p(m + P_1)/(m + P_1 + P_2)\}, & \text{(Phase 3)} \quad p = m + P_1 + P_2, \ldots, m + P\,. \end{cases}$$

*To whom correspondence should be addressed.

1. Phase 1 begins with $p = m$ and $[L^{(m)}, U^{(m)}]$ as starting values. We carry out a P_1-step search with $a_p = c/p$. Garthwaite and Jones (2009) suggest keeping this phase short by choosing $P_1 = \min(5,000, P/20)$. Phase 1 ends with estimates $[L^{(m+P_1)}, U^{(m+P_1)}]$.
2. Phase 2 begins with $p = m + P_1$ and $[L^{(m+P_1)}, U^{(m+P_1)}]$ as starting values. We carry out a P_2-step search with $a_p = c/(m+P_1)$. Garthwaite and Jones (2009) recommend $P_2 = 14 \times P_1$ as a reasonable choice. Phase 2 ends with estimates $[L^{(m+P_1+P_2)}, U^{(m+P_1+P_2)}]$.
3. Phase 3 begins with $p = m + P_1 + P_2$ and $[L^{(m+P_1+P_2)}, U^{(m+P_1+P_2)}]$ as starting values. We carry out a P_3-step search with $a_p = c/\{p(m + P_1)/(m + P_1 + P_2)\}$. Garthwaite and Jones (2009) suggest keeping this phase long; they found particularly good efficiency for $P_3 \geqslant 200,000 - P_1 - P_2$, i.e. an overall $P \geqslant 200,000$. Phase 3 ends with estimates $[L^{(m+P)}, U^{(m+P)}]$.

Rather than choosing the final updated values $[L^{(m+P)}, U^{(m+P)}]$, the unweighted averages of the final n values are taken as the final CI, i.e. $[\bar{L}, \bar{U}]$ where

$$\bar{L} = \frac{1}{n} \sum_{p=m+P-n+1}^{m+P} L^{(p)} \quad \text{and} \quad \bar{U} = \frac{1}{n} \sum_{p=m+P-n+1}^{m+P} U^{(p)} .$$

Garthwaite and Jones (2009) suggest using $n = P - 2P_1$.

B. Supplementary Figures

(See next page)

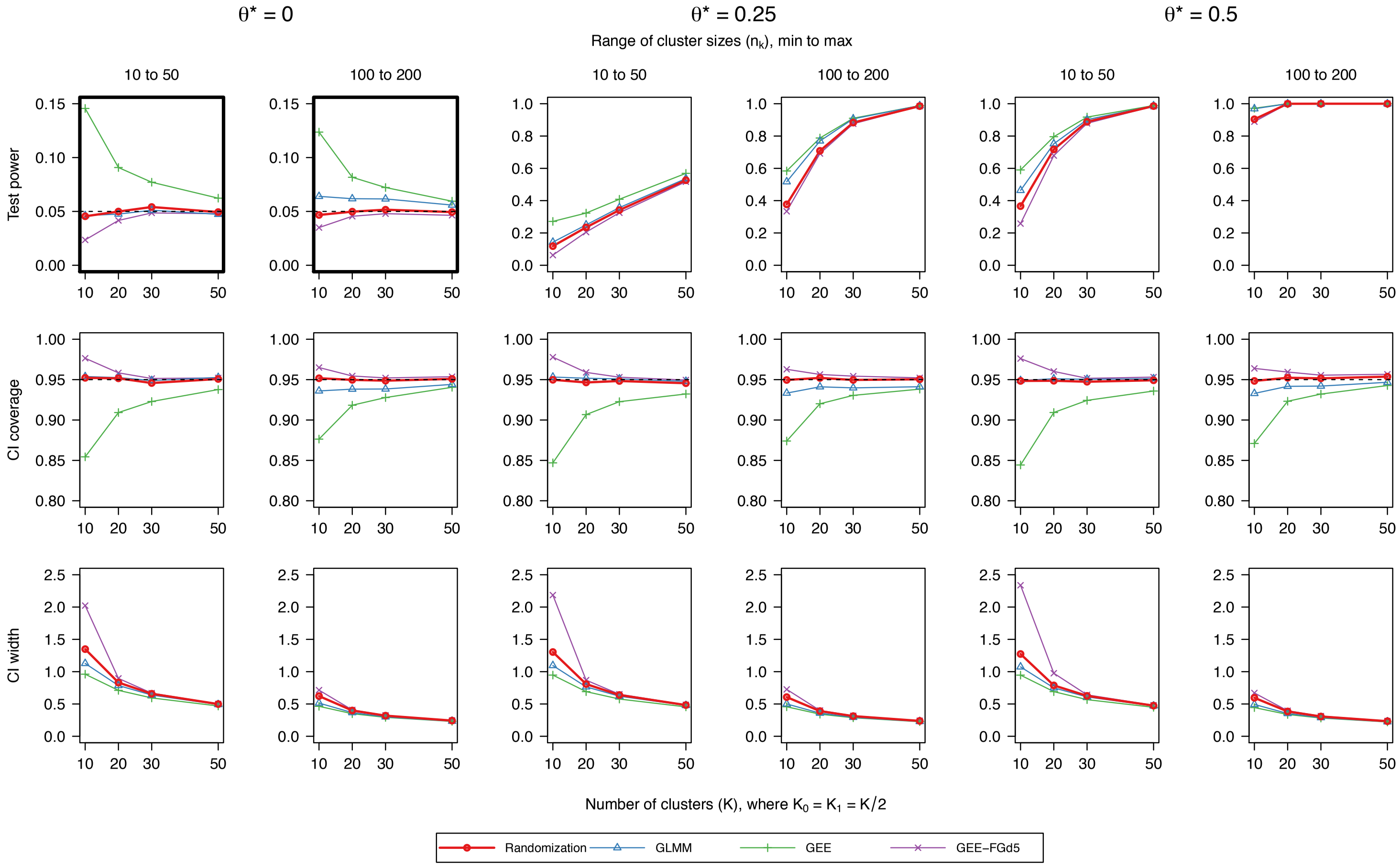

Fig. S1. Simulation results of a parallel CRT with a binary outcome (ICC $\approx$ 0.001). Methods considered include randomization-based inference (Randomization), a logistic generalized linear mixed model fit via maximum likelihood (GLMM), a marginal model fit via a generalized estimating equation (GEE), and a small-sample adjusted GEE (GEE-FGd5). Two bold upper-left plots have a different y-axis range from others in the first row for clarity. This figure appears in color in the electronic version of this article.

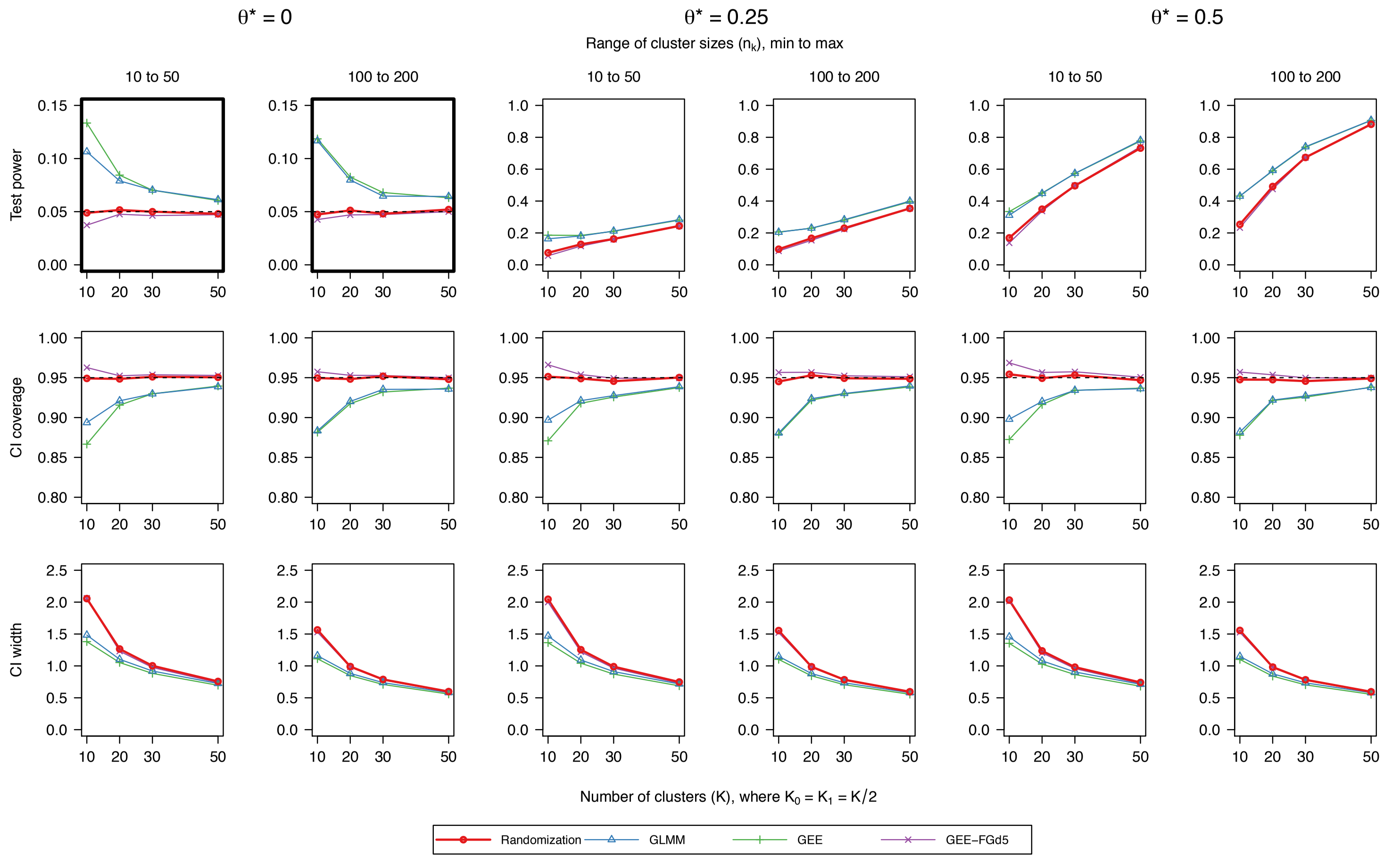

Fig. S2. Simulation results of a parallel CRT with a binary outcome (ICC ≈ 0.05). Methods considered include randomization-based inference (Randomization), a logistic generalized linear mixed model fit via maximum likelihood (GLMM), a marginal model fit via a generalized estimating equation (GEE), and a small-sample adjusted GEE (GEE-FGd5). Two bold upper-left plots have a different y-axis range from others in the first row for clarity. This figure appears in color in the electronic version of this article.

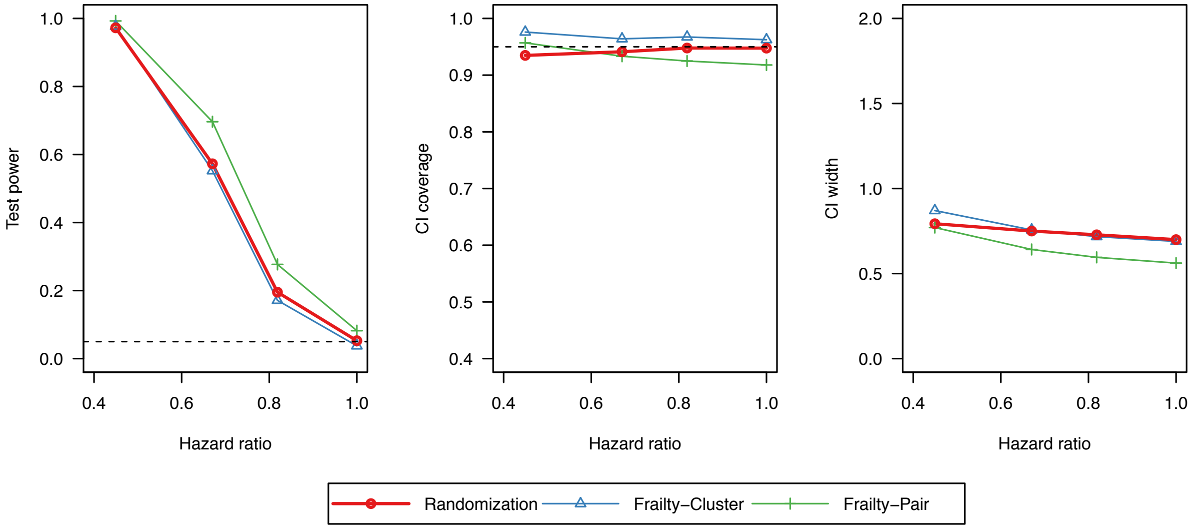

Fig. S3. Simulation results of a pair-matched CRT with an interval-censored time-to-event outcome ($\sigma = 0.2$). Methods considered included randomization-based inference (Randomization), a Weibull model with a frailty term for community (Frailty-Cluster), and with a frailty term for pair (Frailty-Pair). This figure appears in color in the electronic version of this article.

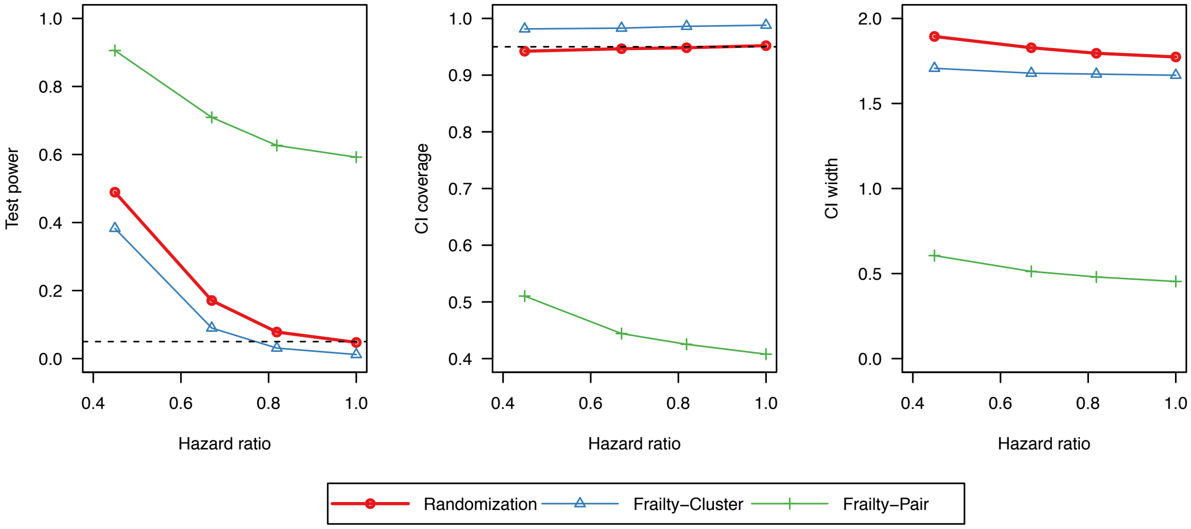

Fig. S4. Simulation results of a pair-matched CRT with an interval-censored time-to-event outcome ($\sigma = 0.8$). Methods considered included randomization-based inference (Randomization), a Weibull model with a frailty term for community (Frailty-Cluster), and with a frailty term for pair (Frailty-Pair). This figure appears in color in the electronic version of this article.

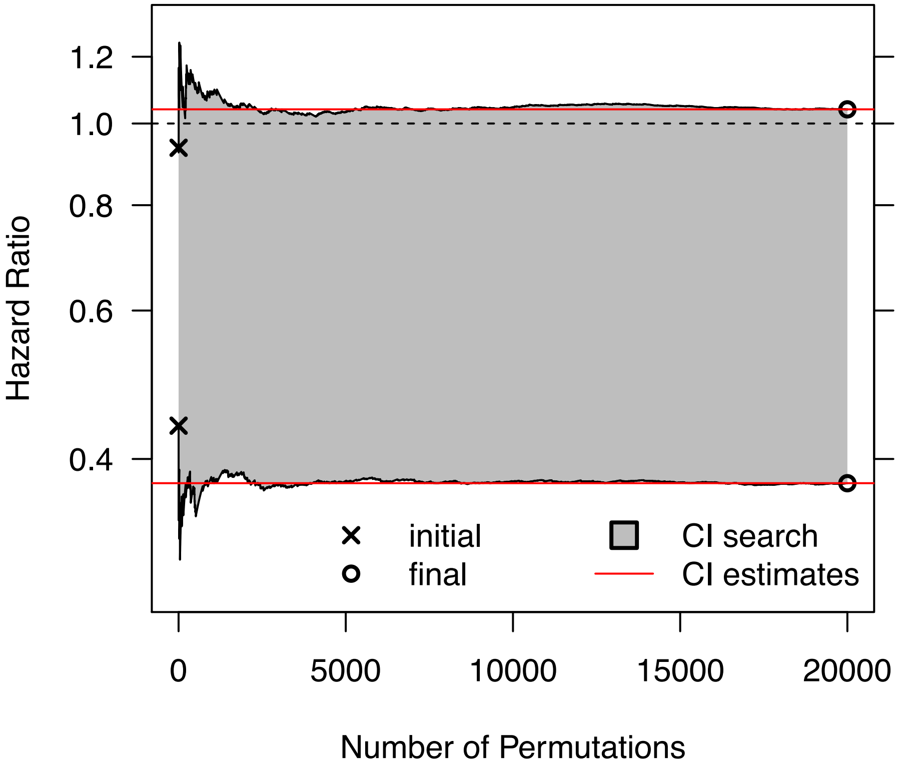

Fig. S5. Efficient search for a randomization-based 95% confidence interval (CI) for the intervention effect in the Botswana Combination Prevention Project, using the procedure adapted from Garthwaite (1996). This figure appears in color in the electronic version of this article.

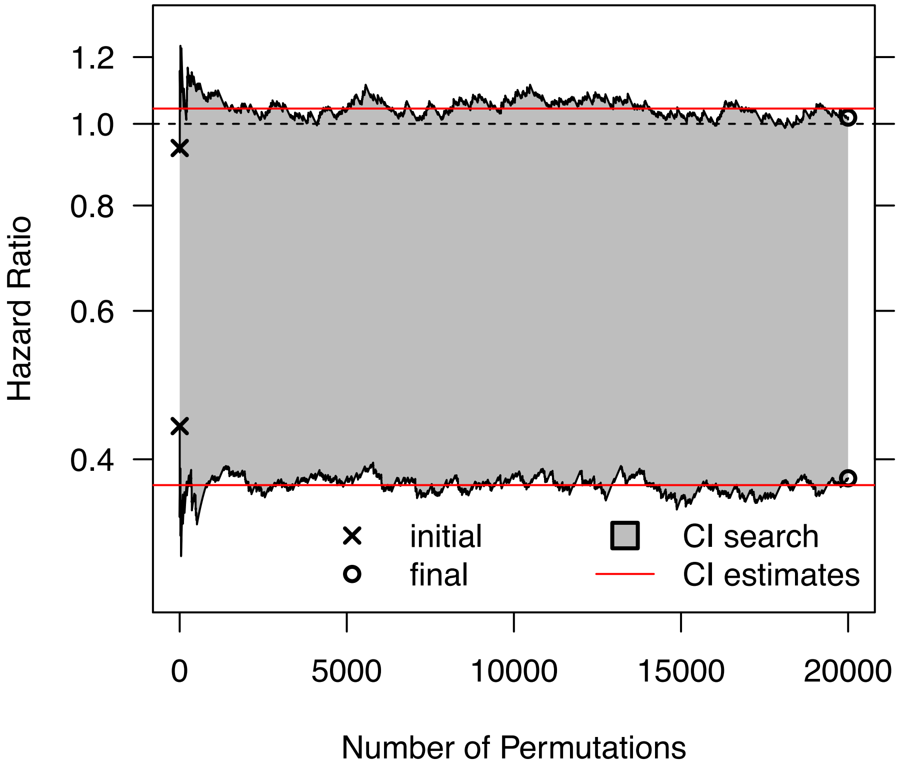

Fig. S6. Efficient search for a randomization-based 95% confidence interval (CI) for the intervention effect in the Botswana Combination Prevention Project, using the procedure adapted from Garthwaite and Jones (2009). This figure appears in color in the electronic version of this article.

C. Supplementary Tables

Table S1. *Average computation time (in seconds) to calculate a single randomization-based CI with $P = 5,000$ permutations for the binary simulation setup. K is the number of clusters, n_k is the cluster size, and $E(N)$ is the mean total sample size. Lower and upper bound searches were run in R in parallel across 2 logical CPUs on a MacBookAir8,1 with a 1.6 GHz Intel Core i5 processor, times being averaged over 10 data sets (for laptop) and on the O2 High Performance Compute Cluster, supported by the Research Computing Group at Harvard Medical School, times being averaged over 100 data sets (for cluster).*

K	n_k	$E(N)$	Time on laptop (sec.)	Time on cluster (sec.)
10	10 to 50	300	23	47
20	10 to 50	600	31	44
30	10 to 50	900	39	58
50	10 to 50	1500	55	86
10	100 to 200	1500	55	82
20	100 to 200	3000	99	171
30	100 to 200	4500	139	264
50	100 to 200	7500	229	738

Table S2. *Model convergence among 10,000 independently generated data sets for different interval-censored time-to-event outcome simulation scenarios (i.e. different values of treatment effect θ^* and SD of cluster and pair random effect σ). Methods considered included randomization-based inference (Randomization), a Weibull model with a frailty term for community (Frailty-Cluster), and with a frailty term for pair (Frailty-Pair). Only 2 Frailty-Cluster and 10 Frailty-Pair (each of 120,000) were deemed not to converge based on a CI width greater than 10^{10}; the remainder failed when attempting to fit the frailty model in R. No Randomization CIs met this non-convergence threshold nor failed when fit in R.*

θ^*	σ	Randomization	Frailty-Cluster	Frailty-Pair
0	0.2	10000	9998	10000
0	0.5	10000	8874	9548
0	0.8	10000	9391	9024
-0.2	0.2	10000	9999	10000
-0.2	0.5	10000	8808	9475
-0.2	0.8	10000	9388	9005
-0.4	0.2	10000	9993	9999
-0.4	0.5	10000	8670	9285
-0.4	0.8	10000	9451	9044
-0.8	0.2	10000	9897	9990
-0.8	0.5	10000	8304	8809
-0.8	0.8	10000	9596	9090

Table S3. *Estimated CI of the intervention effect (hazard ratio) in the Botswana Combination Prevention Project, using the P-step search procedure adapted from Garthwaite and Jones (2009).*

Method	95% CI
$P = 20,000$	
Randomization, Marginal	[0.373, 1.043]
Randomization, Pair-Stratified	[0.369, 1.058]
$P = 200,000$	
Randomization, Marginal	[0.378, 1.031]
Randomization, Pair-Stratified	[0.370, 1.046]

References

Garthwaite, P. H. (1996). Confidence intervals from randomization tests. *Biometrics* **52**, 1387–1393.

Garthwaite, P. H. and Jones, M. C. (2009). A stochastic approximation method and its application to confidence intervals. *Journal of Computational and Graphical Statistics* **18**, 184–200.