Title

Blurring cluster randomized trials and observational studies: Two-Stage TMLE for subsampling, missingness, and few independent units

Permalink

https://escholarship.org/uc/item/60z8540s

Journal

Biostatistics, 25(3)

ISSN

1465-4644

Authors

Nugent, Joshua R
Marquez, Carina
Charlebois, Edwin D
et al.

Publication Date

2024-07-01

DOI

10.1093/biostatistics/kxad015

Peer reviewed

Biostatistics, 2024, **25**, 599–616
https://doi.org/10.1093/biostatistics/kxad015
Advance Access publication August 2, 2023

OXFORD

Blurring cluster randomized trials and observational studies: Two-Stage TMLE for subsampling, missingness, and few independent units

Joshua R. Nugent [1,*], Carina Marquez [2], Edwin D. Charlebois[3], Rachel Abbott[2], Laura B. Balzer [4] for the SEARCH COLLABORATION

[1]Division of Research, Kaiser Permanente Northern California, 2000 Broadway, Oakland, CA 94612, USA
[2]Division of HIV, Infectious Diseases, and Global Medicine, University of California, 1001 Potrero Avenue, San Francisco, CA 94110, USA
[3]Center for AIDS Prevention Studies, University of California, 550 16th Street, San Francisco, CA 94158, USA
[4]Division of Biostatistics, School of Public Health, University of California, 2121 Berkeley Way, Berkeley, CA 94720, USA

[*]To whom correspondence should be addressed: Joshua R. Nugent. Email: joshuanugent@gmail.com

SUMMARY

Cluster randomized trials (CRTs) often enroll large numbers of participants; yet due to resource constraints, only a subset of participants may be selected for outcome assessment, and those sampled may not be representative of all cluster members. Missing data also present a challenge: if sampled individuals with measured outcomes are dissimilar from those with missing outcomes, unadjusted estimates of arm-specific endpoints and the intervention effect may be biased. Further, CRTs often enroll and randomize few clusters, limiting statistical power and raising concerns about finite sample performance. Motivated by SEARCH-TB, a CRT aimed at reducing incident tuberculosis infection, we demonstrate interlocking methods to handle these challenges. First, we extend Two-Stage targeted minimum loss-based estimation to account for three sources of missingness: (i) subsampling; (ii) measurement of baseline status among those sampled; and (iii) measurement of final status among those in the incidence cohort (persons known to be at risk at baseline). Second, we critically evaluate the assumptions under which subunits of the cluster can be considered the conditionally independent unit, improving precision and statistical power but also causing the CRT to behave like an observational study. Our application to SEARCH-TB highlights the real-world impact of different assumptions on measurement and dependence; estimates relying on unrealistic assumptions suggested the intervention increased the incidence of TB infection by 18% (risk ratio [RR] = 1.18, 95% confidence interval [CI]: 0.85–1.63), while estimates accounting for the sampling scheme, missingness, and within community dependence found the intervention decreased the incident TB by 27% (RR = 0.73, 95% CI: 0.57–0.92).

KEYWORDS: Cluster randomized trials (CRTs); Double robustness; Efficiency; Group randomized trials; Hierarchical data; Missing data; Multi-level data; Super Learner; Two-Stage targeted minimum loss-based estimation (TMLE)

Received: August 30, 2023. **Revised:** May 15, 2023. **Accepted:** July 2, 2023

1. Introduction

In randomized controlled trials, the intervention is sometimes randomized to groups of participants rather than to individuals (Hayes and Moulton, 2009; Campbell and Walters, 2014; Donner and Klar, 2010; Eldridge and Kerry, 2012). For example, it would be impractical to evaluate a new teaching method by randomizing students within classrooms, but much more feasible if classrooms were randomized. In group or cluster randomized trials (CRTs), correlation between the outcomes within a cluster may arise due to shared environmental factors, shared exposure to the intervention (or control), and interactions between individuals within a cluster. This dependence violates the common regression assumption that all observations are independent and identically distributed (i.i.d.), complicating statistical estimation and inference.

A number of well-established methods can account for the dependence of observations in a cluster (Liang and Zeger, 1986; Fitzmaurice *and others*, 2012; Hayes and Moulton, 2009). However, not all methods can address practical challenges arising in CRTs. First, outcomes may not be measured on all participants in each cluster. This could occur by design, for example, if measurement of a rare or expensive outcome only occurred in a subsample of participants. Failing to adjust for sampling can result in biased point estimates and misleading inference (Horvitz and Thompson, 1952; Robins, 1986; van der Laan and Rose, 2011). Additionally, incomplete ascertainment of outcomes among all (or the selected subset of) participants can bias results if the outcomes are not missing completely at random (MCAR) (Rubin, 1976; Robins *and others*, 1995). Individuals whose outcomes are not measured are likely different than those who were fully observed; for example, students who are absent on an exam day may be systematically different than those present. If this systematic missingness is influenced by the intervention (e.g., a new teaching technique improves motivation and attendance, influencing exam scores and the probability of measurement), the risk of bias is even larger. This is a common problem: a recent review found that missing data were present in 93% of CRTs, 55% of which simply performed a complete-case analysis (Fiero *and others*, 2016).

Second, resource constraints often limit the number of clusters in CRTs. Indeed, a review of 100 CRTs found 37% with fewer than 20 clusters (Kahan *and others*, 2016) and another review of 100 CRTs found a median of 33 clusters (Selvaraj and Prasad, 2013). Further, in CRTs with many clusters, key subgroup analyses might be conducted within strata defined by cluster-level covariates (e.g., region), limiting the number of randomized units included in that analysis. As the number of clusters shrinks, chance imbalance on covariates that influence the outcome becomes more likely. Accounting for these covariates and other outcome predictors can increase precision (e.g., Fisher, 1932; Tsiatis *and others*, 2008; Moore and van der Laan, 2009; Hayes and Moulton, 2009; Benitez *and others*, 2023). However, in analyses with few clusters, including too many covariates can lead to overfitting, and it is often not clear which covariates to select for optimal performance (Balzer *and others*, 2016b).

Third, statistical inference often relies on (i) tests with known finite sample properties that may be inefficient or (ii) the asymptotic behavior of estimators that may not hold in CRT analyses with a limited number of clusters. For example, generalized estimating equations (GEE) and generalized linear mixed models (GLMMs) are two common approaches for analyzing CRTs (Laird and Ware, 1982; Liang and Zeger, 1986); both rely on having a "sufficient" number of clusters. The exact recommendation varies, with some suggesting GEE can be used with as few as 10 clusters (Pan and Wall, 2002), while others suggest that these approaches (without small-sample corrections) should be avoided without 30 or more clusters (Kreft, 1998; Hayes and Moulton, 2009; Murray *and others*, 2018). Altogether, inference based on a small number of clusters may be unreliable, creating conservative or anticonservative confidence interval coverage depending on the situation (Leyrat *and others*, 2018). For an overview and comparison of methods for CRT analysis, we refer the reader to Hayes and Moulton (2009) and Benitez *and others* (2023).

Here, we address these challenges by combining *Two-Stage targeted minimum loss-based estimation* (TMLE) to account for subsampling and missing individual-level outcomes (Balzer *and others*, 2021) with carefully considered *conditional independence assumptions* to address limited numbers of clusters (van der Laan *and others*, 2013). The novel contributions of this work include the following. First, we extend Two-Stage TMLE to handle differential measurement of an outcome among a closed cohort, where cohort membership is defined by subsampling and also subject to differential measurement at baseline. Second, we detail the assumptions required to increase the effective sample size by considering a subunit of the cluster to be the conditionally independent unit; this process results in the CRT behaving like an observational study. As a consequence, we extend the prior asymptotic results and practical implementation of Two-Stage TMLE for this psuedo-observational setting. Additionally, we discuss how our approach relates to assumptions commonly made in multilevel observational studies, where, for example, individuals are nested in neighborhoods and substantial interactions occur within and across those neighborhoods (Oakes, 2004; Sobel, 2006). Finally, we demonstrate the real-life consequences of various analytic choices, using real-world data from the SEARCH-TB study.

Briefly, SEARCH-TB sought to evaluate the population-level effect of universal HIV test-and-treat on incident tuberculosis (TB) infection in rural Uganda. SEARCH-TB was a substudy of the SEARCH trial, a 32-community CRT (NCT01864603) (Havlir *and others*, 2019). Intervention communities received annual, population-based HIV testing with universal treatment eligibility and patient-centered care delivery. Control communities received population-based testing at baseline with treatment eligibility according to Ministry of Health guidelines. Given logistical and financial constraints, detailed below, assessment of incident TB infection was limited to nine communities, within which a subsample of participants was selected based on the HIV status of their household. Multiple visits were made to selected households to administer sociodemographic surveys and tuberculin skin tests (TSTs) to persons aged 5 years and older. The substudy participants who were TST-negative at baseline formed a closed cohort, on whom follow-up TSTs were attempted 1 year later. The primary outcome of the substudy was the 1-year incidence of TB infection. The applied results have been previously presented (Marquez *and others*, 2022); here, we focus on the causal and statistical methods to account for purposefully differential sampling, potentially differential outcome measurement, and few independent units. Full discussion of the application is given in Section 4; we now present our analytic approach more generally.

2. TWO-STAGE TMLE FOR SAMPLING AND MISSING OUTCOMES

In CRTs, "two-stage" approaches first estimate a cluster-level endpoint and then use those estimates to evaluate the intervention effect (Hayes and Moulton, 2009; Murray *and others*, 2018). As detailed in Benitez *and others* (2023), such approaches can be combined with weighting schemes to estimate cluster-level or individual-level effects on any scale. In particular, Two-Stage TMLE was developed to reduce bias and improve efficiency of CRTs by optimally adjusting for baseline cluster-level covariates, *after* controlling for missingness on individual-level outcomes (Balzer *and others*, 2021). In Stage 1, we identify and estimate a cluster-level endpoint, accounting for potentially differential measurement of individual-level outcomes. To do so, we (i) define a cluster-level counterfactual parameter as a summary of the individual-level counterfactual outcomes of the cluster members, (ii) assess identifiability of that causal parameter, and then (iii) estimate the corresponding statistical parameter in each cluster separately. In Stage 2, we use the resulting endpoint estimates from each cluster to evaluate the intervention effect, optimally adjusting for cluster-level covariates to increase precision. Two-Stage TMLE compares favorably to competing CRT methods, especially when there are post-baseline causes of missingness (Balzer *and others*, 2021). We now extend the approach to account for subsampling and missingness at both baseline and follow-up. In Section 3.2, we further extend the method to support conditional independence assumptions commonly made in observational epidemiology.

2.1. Stage 1: Identifying and estimating the cluster-level endpoint

When the individual-level outcomes are not MCAR, estimating the cluster-specific endpoint with the simple mean among those measured can create several hazards. First, failing to account for over-sampling of certain subgroups and under-sampling of others can bias estimates for the population of interest. Second, in longitudinal studies, failing to account for incomplete measurement of baseline status can skew estimates of baseline prevalence and estimates of intervention effectiveness. As an extreme example, suppose only participants at very low risk of the outcome were tested at baseline; then estimates of baseline prevalence would be biased downwards, and the resulting incidence cohort would be a poor representation of the population at risk. Likewise, failing to account for incomplete measurement of final endpoint status among the longitudinal cohort can also bias estimates of incidence and intervention effectiveness. As another extreme example, suppose all high-risk cohort members did not have their endpoint measured; then cluster-level estimates of incidence would be biased downwards. If missingness is present at both baseline and follow-up, these biases could compound. Further, if missingness is differential by arm—say, the high-risk participants were more likely to be measured at follow-up in the intervention arm—the potential for bias is even greater.

In SEARCH-TB, our motivating study, all of these dangers were present. The subsample was enriched for persons with HIV; measurement of baseline TB status was potentially differential among those sampled, and measurement of incident TB infection was also potentially differential among participants who were TST-negative at baseline. In the following subsection, we discuss our definition of the cluster-level endpoint and describe methods for estimating it, along with relevant assumptions.

2.1.1. *Notation.*

Throughout, we denote cluster-level quantities with superscript c and underlying (possibly unmeasured) quantities with an asterisk. For an individual in a given cluster, let E^c represent the cluster-level covariates (e.g., baseline HIV prevalence) and L_0 the set of individual-level covariates (e.g., age). These are either measured prior to intervention implementation or, at minimum, not impacted by the intervention. Let A^c represent whether the cluster was randomized to the intervention ($A^c = 1$) or the control ($A^c = 0$), and S indicate that an individual was sampled for the substudy. Next, define $Y_0^* \in \{0, 1\}$ as a participant's underlying (possibly unmeasured) outcome status at baseline—specifically, $Y_0^* = 1$ if the participant has the outcome (e.g., TB infection) at baseline and 0 if not. Likewise, define Δ_0 as an indicator that their outcome was measured at baseline; hence, Δ_0 is deterministically 0 if the participant was not sampled ($S = 0$) for the substudy. The observed outcome at baseline is defined as $Y_0 = \Delta_0 \times Y_0^*$, equaling 1 if the participant was measured and had the outcome at baseline. Participants known to be at risk at baseline (i.e., those with $\Delta_0 = 1$ and $Y_0 = 0$) form a closed cohort for incidence measurement. Variables Y_1^*, Δ_1, and Y_1 are the follow-up timepoint analogues. Likewise, let L_1 denote post-baseline variables that may be impacted by the intervention A^c and impact the underlying outcome Y_1^* and its measurement Δ_1 at follow-up.

Altogether, the observed data for a participant are $O = (E^c, L_0, A^c, S, \Delta_0, Y_0, L_1, \Delta_1, Y_1)$. Recall that Stage 1 of our approach involves defining and estimating an endpoint in each cluster separately. Therefore, we can simplify the participant-level data to $O = (L_0, S, \Delta_0, Y_0, L_1, \Delta_1, Y_1)$, because the cluster-level covariates E^c and cluster-level exposure A^c are shared by all members of a given cluster (Balzer *and others*, 2021). A simplified directed acyclic graph showing the relationships between the individual-level variables is shown in Figure 1.

2.1.2. *Definition and identification of the cluster-level causal parameter.*

In Stage 1, we focus on the underlying proportion of cluster members with the outcome at follow-up among those at risk at baseline:

$$\mathbb{P}(Y_1^* = 1 \mid Y_0^* = 0) = \frac{\mathbb{P}(Y_1^* = 1, Y_0^* = 0)}{\mathbb{P}(Y_0^* = 0)} = \frac{\mathbb{P}(Y_1^* = 1, Y_0^* = 0)}{1 - \mathbb{P}(Y_0^* = 1)}. \tag{2.1}$$

This is equivalent to the counterfactual incidence of the outcome under the following hypothetical interventions. First, to ensure outcome ascertainment at baseline, we would include all cluster

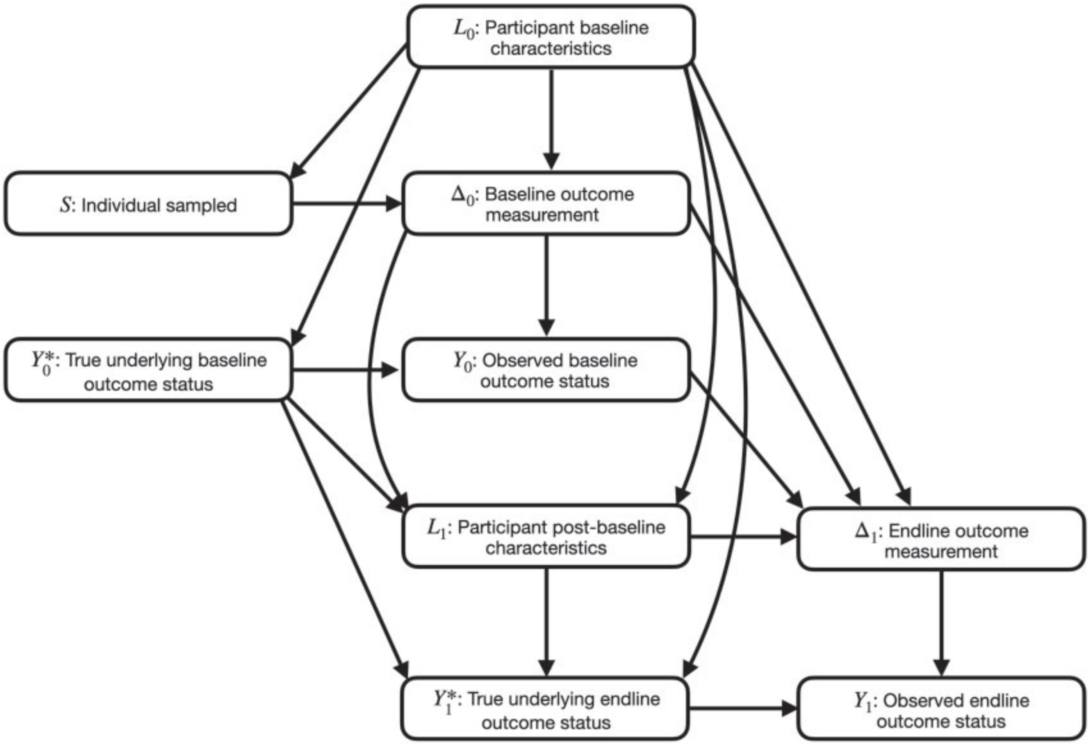
L_0: Participant baseline characteristics
S: Individual sampled
Δ_0: Baseline outcome measurement
Y_0^*: True underlying baseline outcome status
Y_0: Observed baseline outcome status
L_1: Participant post-baseline characteristics
Δ_1: Endline outcome measurement
Y_1^*: True underlying endline outcome status
Y_1: Observed endline outcome status

Figure 1. A simplified causal model to illustrate the relationships between individual-level variables within a cluster in Stage 1. For ease of presentation, the graph is shown without any dependence between unmeasured variables, which are omitted.

members in the study (i.e., "set" $S = 1$) and measure all participants' outcomes (i.e., "set" $\Delta_0 = 1$). Second, to ensure follow-up outcome ascertainment among members of the incidence cohort, we consider a dynamic intervention to "set" $\Delta_1 = 1$ among those at risk at baseline ($Y_0 = 0, \Delta_0 = 1$) (Hernán *and others*, 2006; van der Laan and Petersen, 2007; Robins *and others*, 2008). We now briefly discuss the assumptions needed to express this causal parameter (2.1) as a statistical parameter (i.e., function) of the observed data distribution. The plausibility of the identification assumptions in our motivating example is discussed in Section 4, and further details are in the supplementary material available at Biostatistics online.

For ease of presentation, we reparameterize the denominator of (2.1) as one minus the counterfactual outcome prevalence at baseline: $\mathbb{P}(Y_0^* = 0) = 1 - \mathbb{P}(Y_0^* = 1)$. Under the following assumptions, the latter is identified as the baseline prevalence of the observed outcome, adjusted for differences between participants with measured versus missing outcomes: $\psi^c_{\text{den}} \equiv \mathbb{E}\{\mathbb{E}(Y_0 \mid \Delta_0 = 1, S = 1, L_0)\}$, where superscript c is used to emphasize this statistical parameter is shared by all cluster members. To establish equivalence between $\mathbb{P}(Y_0^* = 1)$ and ψ^c_{den}, we need that subsampling is done randomly within values of L_0 *and* that the only common causes of the outcome and its measurement (among those sampled) are also captured in L_0. This equivalent to assuming baseline outcome status is missing-at-random (MAR): $Y_0^* \perp\!\!\!\perp S \mid L_0$ and $Y_0^* \perp\!\!\!\perp \Delta_0 \mid S = 1, L_0$. Additionally, we need a positivity assumption; subsampling and baseline measurement (among those sampled) is possible, regardless of L_0 values: $\mathbb{P}(S = 1 \mid L_0 = l_0) > 0$ and $\mathbb{P}(\Delta_0 = 1 \mid S = 1, L_0 = l_0) > 0$ for all possible values $l_0 \in L_0$.

Identification of the counterfactual proportion of cluster members who have the outcome at follow-up and are at risk at baseline $\mathbb{P}(Y_1^* = 1, Y_0^* = 0)$ is also possible under two common assumptions. First, the sequential randomization assumption (Robins, 1986) requires that at each timepoint, the MAR assumption holds conditionally on (a subset of) the measured past. This assumption can be evaluated graphically with the sequential backdoor criterion (Pearl, 2009). Second, the corresponding positivity assumption requires a positive probability of measurement at each timepoint, regardless covariate history. Under these assumptions, the numerator of the causal parameter (2.1) can be identified as $\psi^c_{\text{num}} \equiv \mathbb{E}\left[\mathbb{E}\{\mathbb{E}(Y_1 \mid \Delta_1 = 1, L_1, Y_0 = 0, \Delta_0 = 1, S = 1, L_0) \mid \Delta_0 = 1, S = 1, L_0\}\right]$. This is the longitudinal G-computational formula expressed in terms of iterated conditional expectations (Bang and Robins, 2005; van der Laan and Gruber, 2012).

Altogether, our Stage 1 statistical parameter is given by

$$Y^c \equiv \frac{\psi^c_{\text{num}}}{1 - \psi^c_{\text{den}}} = \frac{\mathbb{E}\left[\mathbb{E}\{\mathbb{E}(Y_1 \mid \Delta_1 = 1, L_1, Y_0 = 0, \Delta_0 = 1, S = 1, L_0) \mid \Delta_0 = 1, S = 1, L_0\}\right]}{1 - \mathbb{E}\{\mathbb{E}(Y_0 \mid \Delta_0 = 1, S = 1, L_0)\}}. \tag{2.2}$$

We use Y^c to emphasize this parameter is shared by all cluster members and is a summary measure of the individual-level data within that cluster. Here, Y^c is a complex summary function, but can, nevertheless, be interpreted as the incidence of the outcome, after adjusting for sampling and differential measurement at both baseline and follow-up. Under the identification assumptions, Y^c would equal the counterfactual outcome incidence if there were complete sampling and no missingness: $\mathbb{P}(Y^*_1 = 1 | Y^*_0 = 0)$.

2.1.3. *Estimating the cluster-level statistical parameter.*

Several options exist to estimate the statistical parameters of the denominator ψ^c_{den}, numerator ψ^c_{num}, and, thus, the cluster-level endpoint Y^c. All approaches are implemented in each cluster separately, allowing the relationships between the individual-level covariates, sampling, measurement, and outcomes to vary by cluster *and* naturally accounting for cluster-level variables (E^c, A^c) (Balzer *and others*, 2021). If the adjustment variables (L_0, L_1) are discrete and low-dimensional, we could implement a nonparametric stratification-based approach to estimate the iterated conditional expectations in G-computation or measurement mechanism in inverse probability weighting (Horvitz and Thompson, 1952; Robins, 1986; Bang and Robins, 2005).

However, when the adjustment variables are continuous and/or moderate-to-high dimensional, machine learning can be applied to avoid unsubstantiated modeling assumptions and flexibly learn complex relationships in the data. To support valid statistical inference, machine learning should be incorporated in doubly robust estimators (a.k.a, double/debiased machine learning methods), such as TMLE (van der Laan and Rose, 2011; Díaz, 2019). With respect to our missing data problem (i.e., estimation of Y^c), doubly robust estimators enjoy the following properties: asymptotic linearity under reasonable regularity conditions; consistency if either the iterated conditional expectations or the measurement mechanism is consistently estimated, and efficiency if both are consistently estimated at fast enough rates. As a substitution estimator, TMLE is often preferable to other approaches, especially under data sparsity due to positivity violations or rare outcomes. Implementation of TMLE will vary by parameter and is detailed in the supplementary material available at *Biostatistics* online for both ψ^c_{num} and ψ^c_{den}. We recommend implementing TMLE with Super Learner (van der Laan *and others*, 2007), an ensemble machine learning method, to improve our chances of having both a consistent and efficient estimator.

We obtain a point estimate of the endpoint in each cluster as $\hat{Y}^c = \hat{\psi}^c_{\text{num}}/(1 - \hat{\psi}^c_{\text{den}})$. Then these estimated cluster-level endpoints $\hat{Y}^c_i$ for $i = \{1, \ldots, N\}$ are used to evaluate the intervention effect in Stage 2.

2.2. Stage 2: Definition, estimation, and inference for the treatment effect

Recall our goal of evaluating the intervention effect in a CRT. Let $Y^c(a^c) = \mathbb{P}(Y^*_1(a^c) = 1 \mid Y^*_0(a^c) = 0)$ denote the counterfactual outcome incidence under an additional intervention to "set" $A^c = a^c$. Since the cluster-level treatment is randomized, the following identification conditions hold by design: $Y^c(a^c) \perp\!\!\!\perp A^c$ and $0 < \mathbb{P}(A^c = 1) < 1$. Additionally, since we have already dealt with individual-level sampling and missingness in Stage 1, we can trivially identify summaries of these cluster-level counterfactuals (Balzer *and others*, 2021). Suppose, for example, we are interested in the effect for a population of clusters; then the expected counterfactual outcome $\mathbb{E}[Y^c(a^c)]$ equals the expected outcome among those receiving the exposure of interest $\mathbb{E}(Y^c|A^c)$. Our approach easily accommodates other effects, such conditional or sample effects; however, we focus on population effects throughout this manuscript for demonstration.

To gain efficiency, we incorporate covariate adjustment. Let $\phi^c(a^c) = \mathbb{E}\{\mathbb{E}(Y^c|A^c = a^c, E^c)\}$, where E^c are the baseline covariates, including those measured directly at the cluster-level (e.g., urban vs. rural) and/or aggregates of individual-level covariates L_0 (e.g., HIV prevalence). $\phi^c(a^c)$ is a cluster-level analog of the G-computation identifiability result (Robins, 1986; Balzer *and others*, 2016a, 2019). If the Stage 1 identifiability assumptions hold, contrasts of $\phi^c(1)$ and $\phi^c(0)$ can be interpreted as the population-level intervention effects. If not, contrasts of $\phi^c(1)$ and $\phi^c(0)$ are interpreted statistically as associations of the cluster-level intervention with the incidence of the outcome, after controlling for subsampling and missingness at the individual level.

We now consider how to optimally estimate the Stage 2 statistical parameter ϕ^c, defined as a contrast between $\phi^c(1)$ and $\phi^c(0)$. For example, on the relative scale, $\phi^c = \phi^c(1) \div \phi^c(0)$. In Stage 2, our observed data are at the cluster level: $O^c = (E^c, A^c, \hat{Y}^c)$, where $\hat{Y}^c$ is the cluster-level endpoint estimated in Stage 1.

Using these data, Stage 2 estimation can proceed by implementing a cluster-level analysis, such a G-computation, inverse probability weighting, or TMLE. The key challenge to Stage 2 is *a priori* specification of the optimal adjustment set — which variables and what functional form. One solution to this challenge is to implement *Adaptive Pre-specification* (APS) within TMLE (Balzer *and others*, 2016b). Briefly, APS prespecifies a candidate set of working generalized linear models (GLMs) for the cluster-level outcome regression $\mathbb{E}(\hat{Y}^c|A^c, E^c)$ and for the cluster-level propensity score $\mathbb{P}(A^c = 1|E^c)$ and, then, chooses the combination that minimizes the cross-validated variance estimate for the TMLE of the target parameter. Finite sample simulations and real-data applications have demonstrated substantial precision gains over alternative approaches (Balzer *and others*, 2016b, 2021; Benitez *and others*, 2023).

Under conditions detailed in Balzer *and others* (2021), the Two-Stage TMLE $\hat{\phi}^c$ will be an asymptotically linear estimator of the target effect ϕ^c, such that $\hat{\phi}^c - \phi^c = 1/N \sum_{i=1}^{N} D_i^c + R_N$ with D_i^c as the influence curve (function) for the ith cluster and $R_N = o_p(1/\sqrt{N})$ as the remainder term (van der Vaart, 1998). In particular, we need the contributions from Stage 1 estimation to the remainder term R_N to be essentially zero. Practically, this means we should not bet on bias cancellations when defining or estimating the cluster-level endpoint $Y^c = \psi^c_{\text{num}}/(1 - \psi^c_{\text{den}})$. Indeed, biased estimators of the cluster-level endpoints can result in biased estimates of and misleading inference for the intervention effect. Instead, we recommend using TMLE, incorporating machine learning, to flexibly estimate the cluster-level endpoint Y_i^c for $i = \{1, \ldots, N\}$ in Stage 1. Additionally, two-stage approaches are most effective when the cluster size is relatively large, allowing for adaptive and well-supported estimation of the cluster-level endpoints. The regularity conditions required of Stage 2 estimators of the cluster-level outcome regression and known propensity score hold by design, when using APS to select from working GLMs in TMLE. As discussed next, however, the conditions on the Stage 2 estimators will change if we make alternative identification assumptions in Stage 2.

Under the above conditions, Two-Stage TMLE will be normally distributed in the large data limit, allowing for the construction of Wald-type confidence intervals as $\hat{\phi}^c \pm 1.96\hat{\sigma}$, where $\hat{\sigma}^2$ is the sample variance of the estimated cluster-level influence curve $\hat{D}^c$, scaled by sample size N. (The form of the influence curve will depend on the target parameter ϕ^c.) In CRTs with fewer than 40 clusters randomized ($N < 40$), we recommend using the Student's t distribution with $N - 2$ degrees of freedom as a finite sample approximation of the asymptotic normal distribution (Hayes and Moulton, 2009).

3. (RE-)DEFINING THE INDEPENDENT UNIT

A fundamental premise of CRTs is that outcomes are dependent within a cluster. Sources of dependence could include shared cluster-level factors, including the intervention, as well as social interactions between participants within a cluster. Instead, clusters are assumed to be independent, providing the basis for statistical inference, as described in the prior subsection. However, CRTs tend to randomize few clusters, limiting statistical power. For example, while its parent trial randomized 32 communities, measurement of incident TB infection in SEARCH-TB occurred in only nine communities in Uganda. Even if a given CRT has many clusters, subgroup analyses to understand effect heterogeneity may be conducted among limited numbers of clusters. The extreme case of randomizing to only two clusters, a *de facto* observational study, was covered in depth by van der Laan *and others* (2013).

In this section, our goals are to (i) define a hierarchical causal model, reflecting the data-generating process for a CRT, (ii) detail the assumptions needed to consider a subunit of the cluster to be the conditionally independent unit, and (iii) present the consequences of these assumptions for statistical estimation and inference with Two-Stage TMLE. The level of clustering and, thereby, the definition of "subunit" will vary by setting. In SEARCH-TB, for example, individuals are nested within households, villages, parishes, and communities. Under different assumptions, explicitly stated below, any level of partitioning of the cluster could be treated as the conditionally independent unit.

For simplicity, we focus on CRTs with three layers of clustering: individuals are grouped into subcluster "partitions", indexed by $j = \{1, \ldots, J\}$, and these partitions are grouped into a cluster, which remain the unit of randomization. As before, we denote cluster-level variables with superscript c. Now, denote partition-level variables with superscript p. Recall E^c is the set of cluster-level characteristics; these are sometimes called "environmental" factors, because they represent the shared environment of individuals in a given cluster (van der Laan *and others*, 2013). As before, A^c is an indicator of the cluster being randomized to the intervention arm. Now, let W^p_j be the set of baseline covariates for partition j; these could be general characteristics of the partition (e.g., urban vs. rural) as well as aggregates of baseline covariates of individuals from that partition (e.g., HIV prevalence). Likewise, let Y^p_j be the j^{th} partition's endpoint, which is defined analogously to Y^c in Stage 1 (2.2). Specifically, Y^p_j is the incidence of the outcome, after adjusting for sampling and differential measurement among members of partition j. Under the identification assumptions given in Section 2.1.2, Y^p_j would equal the counterfactual incidence of the outcome for partition j if we had complete sampling and no missingness.

3.1. Hierarchical structural causal models

Using the nonparametric structural causal model of Pearl (2009), we now formalize the hierarchical data generating process for a CRT. For ease of presentation, we focus on CRTs with $J = 2$ partitions per cluster; however, our results naturally generalize to other settings.

Figure 2 provides a causal model, assuming independence between clusters and randomization of the cluster-level intervention ($U_{A^c} \perp\!\!\!\perp U_{E^c}, U_{W^p_1}, U_{W^p_2}, U_{Y^p_1}, U_{Y^p_2}$). The structure of the remaining Us may be complex and cluster-specific; for example, the unobserved factors influencing the partition-level outcomes ($U_{Y^p_1}, U_{Y^p_2}$) might be related to unmeasured, environmental factors U_{E^c}. Beyond the unmeasured factors, there are several sources of dependence between partition-level outcomes in this model. For example, the j^{th} partition's outcome Y^p_j may depend on the characteristics of the other W^p_{-j}. This general causal model encodes independence at the cluster-level, not the partition-level — yet.

To treat the subcluster partition as the conditionally independent unit, we need several assumptions to hold, resulting in a more restrictive causal model reflected in Figure 3 (van der Laan *and others*, 2013). First, there is no interference between partitions within a cluster. Second, any effect of the cluster-level covariates E^c on the partition-level outcome Y^p_j is only through

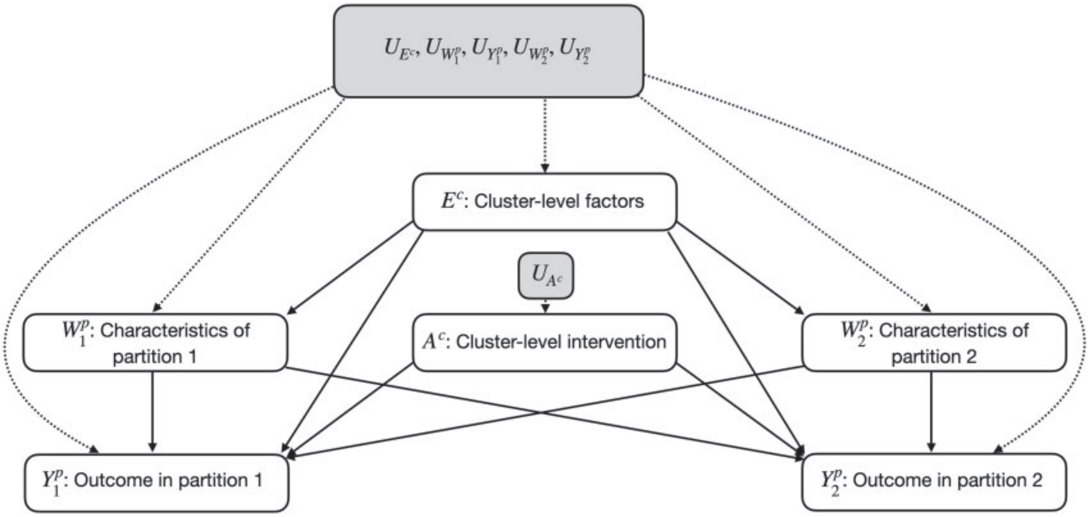
$U_{E^c}, U_{W_1^p}, U_{Y_1^p}, U_{W_2^p}, U_{Y_2^p}$
E^c: Cluster-level factors
U_{A^c}
W_1^p: Characteristics of partition 1
A^c: Cluster-level intervention
W_2^p: Characteristics of partition 2
Y_1^p: Outcome in partition 1
Y_2^p: Outcome in partition 2

Figure 2. A simplified causal model for the data-generating process of a cluster randomized trial with two partitions (i.e., subunits) per cluster. By design in a cluster randomized trial, the unmeasured factors contributing to the cluster-level intervention U_{A^c} are independent of the others. We make no other exclusion restrictions or independence assumptions.

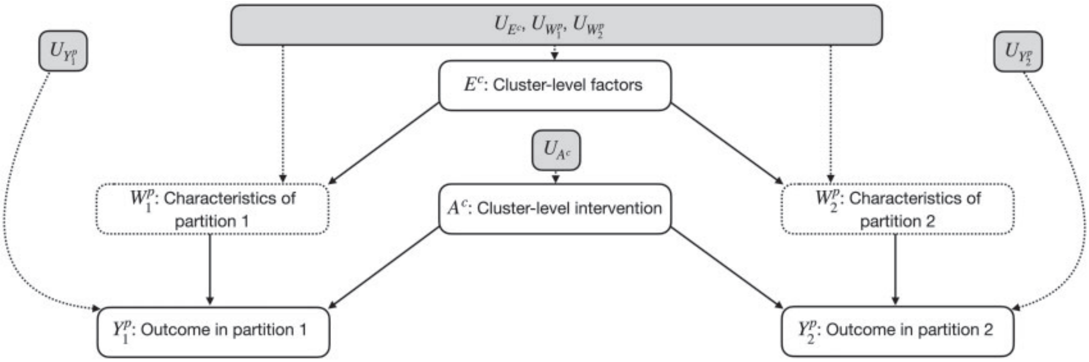
$U_{E^c}, U_{W_1^p}, U_{W_2^p}$
$U_{Y_1^p}$
$U_{Y_2^p}$
E^c: Cluster-level factors
U_{A^c}
W_1^p: Characteristics of partition 1
A^c: Cluster-level intervention
W_2^p: Characteristics of partition 2
Y_1^p: Outcome in partition 1
Y_2^p: Outcome in partition 2

Figure 3. A restricted causal model for the data generating process of a cluster randomized trial with 2 partitions (i.e., subunits) per cluster and under the assumptions needed for the partitions to be conditionally independent. This graph reflects the following exclusion restrictions and independence assumptions: no interference between partitions; no direct effect of the cluster-level covariates E^c on the partition-level outcomes Y^p, and no unmeasured common cause of the partition-level outcomes (U_{Y^p}) and the cluster-level or partition-level covariates (U_{E^c}, U_{W^p}). By design in a cluster randomized trial, the unmeasured factors contributing to the cluster-level intervention (U_{A^c}) are independent of the others.

their effect on j^{th} partition's covariates W_j^p. Finally, there are no unmeasured common causes of partition-level outcomes Y_j^p and the cluster-level or partition-level covariates (E^c, W_j^p). While we additionally need the unmeasured factors contributing to the cluster-level intervention A^c to be independent of the others, this holds by design in CRT. Altogether, these assumptions require there to be no interactions between partitions within a cluster *and* the partition-level covariates W^p are sufficient to block the effects of the cluster-level, environmental factors E^c on the partition-level outcomes Y^p. If these assumptions hold, the partition becomes the conditionally independent unit, increasing the effective sample size, while still allowing for arbitrary dependence *within* each partition.

Whether or not these assumptions are reasonable depends on the study context. To maximize the effective sample size, it might be tempting to define the "partitions" as the individuals in a cluster. However, this approach would entail very strong and possibly unrealistic assumptions, especially in the setting of infectious or contagious outcomes. Instead, if partitions are large subunits of the cluster (e.g., distant neighborhoods in a rural community), these assumptions might be reasonable. Altogether, the assumptions needed to treat the partition as the conditionally independent unit are strong; however, they are commonly evoked in multilevel, observational epidemiology (Oakes, 2004; Sobel, 2006). By explicitly stating them and illustrating them with a causal graph, we aim to empower readers to judge whether they are plausible. Additionally, the design of future studies can be improved by measuring a rich set of covariates to improve the plausibility of these assumptions.

3.2. Estimation and inference with partition-level conditional independence

The assumptions encoded in the restrictive causal model (Figure 3) have important implications for our two-stage estimation approach. Previously, when considering the cluster to be the independent unit, we identified and estimated a cluster-level endpoint Y^c that accounted for subsampling of individuals within that cluster, missingness on baseline outcome status of sampled individuals, and missingness on final outcome status of individuals known to be at risk at baseline. Under the more restrictive model, we now identify and estimate a partition-level endpoint Y^p in Stage 1. Practically, this means that within each partition separately, we use TMLE to estimate $Y^p = \psi^p_{\text{num}}/(1 - \psi^p_{\text{den}})$, as defined in (2.2), and then use the resulting estimates $\hat{Y}^p$ to evaluate the intervention effect in Stage 2.

During effect estimation in Stage 2, we previously adjusted for cluster-level covariates E^c simply to increase precision in a CRT. Now, however, blurring the lines between randomized trials and observational studies *requires* us to adjust for confounders W^p to identify the causal effect and support the conditional independence assumptions. Recall adjustment for the partition-level covariates W^p is required to block the effect of the cluster-level environmental factors E^c, which are no longer included in the adjustment set. Therefore, the Stage 2 statistical estimand is now defined in terms of contrasts of the expected partition-level endpoint, given the cluster-level treatment and partition-level confounders: $\phi^p(a^c) = \mathbb{E}\{\mathbb{E}(Y^p|A^c = a^c, W^p)\}$. For example, on the relative scale, our statistical estimand would be $\phi^p = \phi^p(1) \div \phi^p(0)$. As noted earlier, our approach can target other effects, such as the conditional or sample effects, defined on any scale.)

Importantly, the revised statistical estimand ϕ^p has a subtly different interpretation than the original statistical estimand ϕ^c, which was in terms of the expected cluster-level outcome. If the number of partitions per cluster varies, the value of these two estimands could differ; however, weights can be applied to recover either estimand (Benitez *and others*, 2023). Statistically, ϕ^p can be interpreted as the association of the cluster-level intervention with the incidence of the outcome, after controlling for subsampling and missingness at the individual level and for confounding from environmental factors at the partition level. However, if the Stage 1 identifiability assumptions hold *and* the Stage 2 identifiability assumptions hold, ϕ^p can be interpreted as the population-level intervention effect. The revised Stage 2 statistical estimand ϕ^p could be estimated with a variety of methods. We again recommend TMLE, given its double robustness property, potential for efficiency, and ability to incorporate machine learning while maintaining the basis for valid statistical inference. To implement TMLE for ϕ^p in this setting, we pool together partition-level observations $O^p_k = (W^p_k, A^c_k, \hat{Y}^p_k)$ for the $k = \{1, \ldots, K\}$ partitions in the CRT. Now, $\hat{Y}^p$ represents the estimated partition-level endpoint from Stage 1. Using these data, we implement TMLE at the partition level as if we had a point-treatment observational study (van der Laan and Rose, 2011).

Treating the partition as the conditionally independent unit changes our approach to statistical inference. Specifically, our effective sample size is now K, the number of partitions. However, this comes at the cost of stronger conditions for Two-Stage TMLE $\hat{\phi}^p$ to be asymptotically linear for

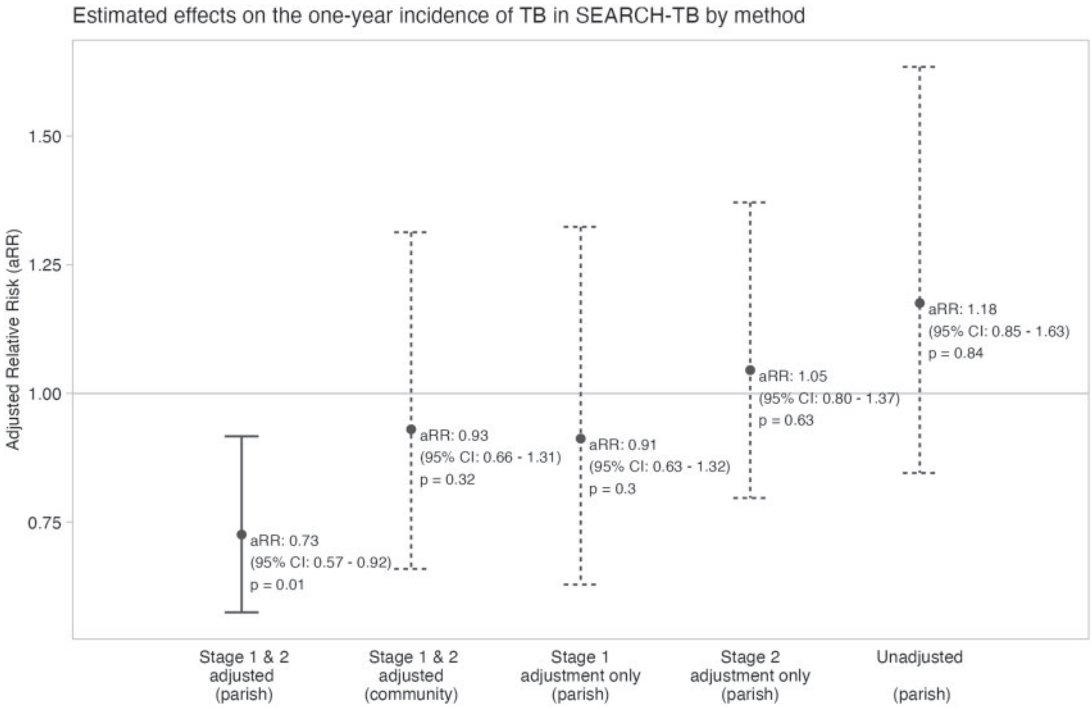

Figure 4. Comparative results under different sets of assumptions using real data from the SEARCH-TB study on incident tuberculosis (TB) infection. The primary analysis, adjusting in both Stage 1 and Stage 2 and considering the parish (a large subunit of the community) to be the conditionally independent unit, is shown first. See Table 1 of the supplementary material available at *Biostatistics* online for additional information.

the target parameter ϕ^p, such that $\hat{\phi}^p - \phi^p = 1/K \sum_{i=1}^{K} D_k^p + R_K$ with D_k^p as the influence curve (function) for the k^{th} partition and $R_K = o_p(1/\sqrt{K})$ as the remainder term. Now, we need the Stage 1 estimators of the partition-level endpoint Y^p to contribute negligibly to the remainder term (Balzer *and others*, 2021). Furthermore, the regularity conditions on effect estimation in Stage 2 do not hold by design. Instead, we need estimators of the partition-level outcome regression $\mathbb{E}(\hat{Y}^p \mid A^c, W^p)$ and partition-level propensity score $\mathbb{P}(A^c \mid W^p)$ to converge to the truth at quick enough rates and avoid overfitting (van der Laan and Rose, 2011). To satisfy these conditions, we again recommend implementing TMLE with Super Learner, considering a diverse set of candidate algorithms, in both Stage 1 and Stage 2.

4. APPLICATION TO THE SEARCH-TB STUDY

An estimated 1.7 billion people, approximately a quarter of the world's population, are infected with TB, and this vast reservoir fuels TB disease and death (Houben and Dodd, 2016; MacPherson *and others*, 2009). Understanding TB transmission dynamics and then implementing effective public health interventions is difficult (Marquez *and others*, 2022). First, transmissions are airborne and likely occur both inside and outside the household. Second, the focus has largely been on active TB (i.e., TB disease), missing the majority of transmission events, which are latent infections. Finally, measurement of latent TB infection through tuberculin skin tests (TSTs) is expensive and imperfect.

Due to resource constraints, evaluation of SEARCH's universal HIV test-and-treat intervention on incident TB infection was conducted through a substudy known as SEARCH-TB

(Marquez *and others*, 2022). This substudy was limited to nine communities in eastern Uganda and 100 randomly sampled households in each community. As previously discussed, househod sampling was enriched for persons with HIV. Among members of the sampled households, latent TB infection was measured via door-to-door placement and reading of TSTs. Incident TB infection was defined as conversion from a negative to positive TST after 1 year of follow-up. Finally, given few randomized clusters, *parishes*, a subunit of the community (analagous to the partitions discussed in Section 3), were considered to be the conditionally independent unit under the assumptions detailed below.

4.1. Stage 1: Identification and estimation of the one-year incidence of TB infection in each partition

We first defined and estimated a partition-level endpoint Y^p, appropriately accounting for subsampling and differential TB ascertainment at the individual level. Of the 17 858 households in the nine study communities, 1435 were sampled, and 688 (47.9%) of the sampled households had at least one adult (aged 15 and up) with HIV. The adult prevalence of HIV in the subsample was 19.6%, a sharp contrast to the prevalence in the region of 3.6% (Havlir *and others*, 2019). Since the risk of TB differs by HIV serostatus (MacPherson *and others*, 2009), ignoring the sampling scheme would bias estimates of TB burden and the intervention effect. However, sampling S was random within household HIV status H. Thus, the following assumptions were satisfied by design: $Y_0^* \perp\!\!\!\perp S \mid H$ and $\mathbb{P}(S=1 \mid H=h) > 0$ for $h \in \{0,1\}$.

Despite up to three visits to the sampled households, including weekends and after hours, TSTs were administered to 4884/8420 (58%) of household members at baseline. Known risk factors for prevalent TB and missingness include age and mobility (Marquez *and others*, 2022). Let W represent these baseline individual-level risk factors. We were willing to assume that for sampled individuals and within values of W, TB prevalence among those with a baseline TST was representative of TB prevalence among those without a baseline TST: $Y_0^* \perp\!\!\!\perp \Delta_0 \mid W, S = 1, H$. Additionally, we assumed that among those sampled, there was a positive probability of administering a TST within all possible values of W. These assumptions, together with the sampling design, allowed for the identification of the counterfactual baseline TB prevalence in each partition $\mathbb{P}(Y_0^*=1)$ as $\psi_{\text{den}}^p = \mathbb{E}\{\mathbb{E}(Y_0 \mid \Delta_0 = 1, W, S = 1, H)\}$.

Among the 4884 participants with known baseline TB status, 3831 (78%) were TST-negative, forming a closed cohort for incidence measurement. As before, despite best efforts, follow-up TST administration was imperfect, with 2425/3831 (63%) of the cohort measured at follow-up. To address potentially differential ascertainment of follow-up status, we considered common risk factors for incident TB infection and its measurement. Given the epidemiology of the region, we again identified age, mobility, and household HIV status as key joint causes of outcomes and missingness. We assumed that within values of these adjustment factors, the risk of incident TB infection among cohort members with a follow-up TST was representative of the risk among cohort members without a follow-up TST. We also assumed a positive probability of receiving a follow-up TST (among the incidence cohort) within all values of (W, H). These assumptions were again supported by the study design, including the repeat visits to households, and allowed for identification of the counterfactual proportion with TB at follow-up but not at baseline $\mathbb{P}(Y_1^* = 1, Y_0^* = 0)$. The corresponding statistical estimand was $\psi_{\text{num}}^p = \mathbb{E}[\mathbb{E}\{\mathbb{E}(Y_1 \mid \Delta_1 = 1, Y_0 = 0, \Delta_0 = 1, W, S = 1, H) \mid \Delta_0 = 1, W, S = 1, H\}]$, which is a simplified version of the ψ_{num}^c parameter from Section 2.1 but without the time-dependent covariates L_1.

For estimation and inference in Stage 1, we stratified on parish, the assumed conditionally independent unit, and estimated ψ_{num}^p and ψ_{den}^p with a participant-level TMLE using Super Learner to combine predictions from main-terms GLM, multivariate adaptive regression splines, and the simple mean. Then for each parish, we obtained estimates of the 1-year incidence of TB infection as $\hat{Y}^p = \hat{\psi}_{\text{num}}^p \div (1 - \hat{\psi}_{\text{den}}^p)$.

4.2. Stage 2: Evaluation of the intervention effect in SEARCH-TB

Next, we used the Stage 1 endpoint estimates $\hat{\psi}_k^p$ for $k = \{1, \ldots, K\}$ to evaluate the effect of the cluster-level intervention in Stage 2. Before doing so, we needed to critically evaluate the assumptions needed to treat the K subcommunity partitions as the conditionally independent unit. Given the following considerations, we immediately eliminated the individual and the household as possible candidates. First, factors influencing TB infection risk include both an individual's susceptibility (e.g., age and HIV status) as well as their level of exposure to TB. Within a household, one member's risk factors could influence their own TB status as well as the TB status of the other household members, especially in settings with poor ventilation and shared sleeping areas. This directly violates the assumption of no interference between individuals within a household. Furthermore, an estimated 80% of TB cases are acquired outside of the household (Martinez *and others*, 2017, 2019) — violating the potential assumption of no interference between households.

Therefore, for the following reasons, we assumed the parish, a large subunit of the community, to be the conditionally independent unit. First, we considered how and where TB is transmitted outside the home in rural Ugandan communities. Prior studies from high-TB burden countries in Sub-Saharan Africa have shown clinics, schools, churches, and workplaces are the areas of high TB risk (Andrews *and others*, 2014). Additionally, prior molecular epidemiologic studies in Uganda have highlighted the role of bars in TB transmission (Chamie *and others*, 2015, 2018). After conducting community mapping and having detailed discussion with the larger Ugandan research team, we concluded these locations are generally shared within a parish, but it was unlikely people would travel between parishes to visit these locations. Therefore, we were willing to assume that there was negligible interference between parishes within a commmunity.

We then considered whether the measured parish-level covariates were sufficient to block the effects of the environmental, community-level factors. First, the role of HIV in fueling the TB epidemic is well established; the biomedical mechanism is via immunosuppression leading to increased susceptibility to infection and reactivation of latent TB infections (Getahun *and others*, 2010). Additionally, the relationship between TB and alcohol has been well established. Globally, an estimated 10% of TB disease is attributable to alcohol use disorder (Rehm *and others*, 2009), and a large systematic review found a 3-fold higher risk of TB disease associated with alcohol use disorder (Lönnroth *and others*, 2008). Our team's prior research in Uganda has also demonstrated a dose–response relationship between levels of alcohol use and latent TB infection (Puryear *and others*, 2021). The underlying mechanisms include alcohol-induced immunosuppression and increased exposure to TB due to time-spent in bars, which are high-TB-risk venues. Altogether, we were willing to assume that the parish-level characteristics of HIV prevalence and prevalence of adults who drink alcohol W^p were sufficient to block the influences of other community-level covariates E^c on the 1-year incidence of TB infection in each parish Y^p. Under these assumptions and with two parishes per community, the effect sample size was $K = 18$. For estimation and inference of the relative effect $\phi^p = \mathbb{E}\{\mathbb{E}(Y^p|A^c = 1, W^p)\} \div \mathbb{E}\{\mathbb{E}(Y^p|A^c = 0, W^p)\}$ in Stage 2, we implemented a parish-level TMLE with Super Learner using the same library of prediction algorithms. Computing code is available at https://github.com/joshua-nugent/search-tb.

4.3. Results of the real-data analysis

The results of the SEARCH substudy on incident TB infection have been previously presented in Marquez *and others* (2022). The primary prespecified analysis, using Two-Stage TMLE with the parishes as the conditionally independent unit, suggested that the universal HIV test-and-treat intervention resulted in a 27% reduction in incident TB infection in eastern Uganda; the adjusted relative risk (aRR) was 0.73 (95% CI: 0.57−0.92; p=0.005).

We now explore the practical impact of varying the identfication assumptions on estimation and inference. The results of our comparison are summarized in Figure 4 and Table 1 in the supplementary material available at *Biostatistics* online. First, we relaxed the assumption that parishes were conditionally independent and, instead, took a more traditional approach treating the randomized unit (i.e., the community) as the independent unit. As expected, when we moved from a parish-level analysis ($K = 18$) to a community-level analysis ($N = 9$), the effect estimate shifted and substantial precision was lost: aRR = 0.93 (95% CI: 0.66−1.31; $p = 0.32$). In this secondary analysis, Stage 1 was implemented analogously to obtain community-level estimates of TB incidence, accounting for sampling and missingness at the individual level. However, Stage 2 effect estimation was done at the community-level with TMLE, using Adaptive Prespecification to select the adjustment covariates to maximize empirical efficiency in the CRT (Balzer *and others*, 2016b).

To further explore the impact of our assumption that parishes were conditionally independent, we conducted a sensitivity analysis where Stage 1 accounted for missingness (as before), but Stage 2 was implemented without adjustment. This approach corresponds to the very strong assumption that the only source of dependence between parishes was the shared community-level intervention A^c. In other words, this analysis assumed no community-level covariates (measured or not) directly or indirectly influenced the incidence of TB infection. Estimates from this approach were again in the similar direction, but even less precise: aRR = 0.91 (95% CI: 0.63−1.32; $p = 0.30$).

Next, we explored the impact our missing data assumptions. Specifically, we conducted a sensitivity analysis where Stage 1 estimates of incidence were unadjusted, but Stage 2 was adjusted (as before). This approach corresponds to the very strong and unreasonable assumption that individual-level outcomes were MCAR. In fact, we know this assumption was violated: the subsample was enriched for persons with HIV, and HIV is a known risk factor for TB. Age and mobility are additional risk factors for TB and for not having a TST administered at baseline or follow-up. Estimates from the approach were markedly different and in the opposite direction of the primary analysis: aRR = 1.05 (95% CI: 0.80–1.37; $p = 0.63$). In other words, conducting a complete-case analysis would lead to the conclusion that the SEARCH intervention *increased* the incidence of TB infection by 5%.

Finally and as an extreme example of strong assumptions on measurement and dependence, we conducted a fully unadjusted analysis. In Stage 1, we estimated the parish-level incidence of TB infection with the raw proportion among those measured. Then in Stage 2, we compared parish-level incidence estimates by arm without further adjustment. This approach is not recommended in practice and suggested the SEARCH intervention *increased* the incidence of TB infection by 18%: aRR = 1.18 (95% CI: 0.85–1.63; $p = 0.84$).

5. DISCUSSION

Cluster randomized trials (CRTs) allow for the rigorous evaluation of interventions delivered at the group-level. Within CRTs, rare or expensive outcomes may only be measured in a subset of clusters and, within those clusters, on a subsample of participants. Missing outcomes among participants is another common issue, which can bias estimates of baseline prevalence, the incidence of the outcome, and the intervention effect. To address these challenges, we extended Two-Stage TMLE to account for subsampling of participants and differential measurement of their outcomes at baseline and at follow-up. Additionally, we detailed the assumptions needed to consider a subcluster partition as the conditionally independent unit. We also extended Two-Stage TMLE to this novel setting, which blurs the lines between CRTs and observational studies. Our application to real-data from SEARCH-TB demonstrated the real-world impact of varying assumptions and analytic choices. For example, ignoring the sampling scheme and assuming the outcomes were missing-completely-at-random reversed the direction of the estimated intervention effect.

When estimating the endpoint in Stage 1 and evaluating the intervention effect in Stage 2, we used TMLE with Super Learner to avoid parametric assumptions and, instead, support efficient estimation in large, semiparametric models. In the absence of missing data, a single-stage approach, such as GLMMs or GEE, could be used to estimate the intervention effect if the effective sample size is sufficiently large. These methods account for the dependence of participants within a partition and can incorporate adjustment for partition-level variables W^p needed to support the independence assumptions. However, when adjusting for covariates, these alternative estimators are often limited in their ability to estimate marginal effects (Benitez *and others*, 2023). For example, when using the logit-link in GLMM and GEE, the conditional odds ratio is estimated (Laird and Ware, 1982; Hubbard *and others*, 2010). Additionally, as previously discussed, even after considering the subcluster partition to be the conditionally independent unit, the effective sample size may still be too small to support use of these approaches without finite sample corrections. Finally and perhaps most importantly, these methods cannot accommodate post-baseline causes of missingness (Balzer *and others*, 2021). Altogether, to handle common analytic challenges in CRTs (e.g., differential missingness and few clusters) and to estimate marginal effects on any scale, we recommend using TMLE, a doubly robust, semi-parametric efficient, substitution estimator, in our two-stage approach.

Nonetheless, our approach does require real assumptions on the missingness mechanism and the dependence structure within a cluster. These assumptions have implications for trial design. First, all the shared causes of missingness and outcomes must be measured. Second, fairly large cluster sizes (or subcluster partition sizes) are needed for stable and consistent estimation of the endpoints in Stage 1. Finally, to support any conditional independence assumptions and improve precision in Stage 2, a rich set of partition-level covariates should be collected. We again emphasize these conditional independence assumptions are commonly made, but less commonly acknowledged, in multilevel observational studies (Oakes, 2004; Sobel, 2006).

In all cases, these assumptions should be carefully considered, transparently stated, and illustrated with a causal graph. As discussed in the real-data example, assuming individuals or households are effectively independent might be unrealistic in many settings. Alternatively, considering larger partitions of the cluster, such as distant neighborhoods, might be more reasonable. While larger partitions weakens the required identification assumptions, fewer (conditionally) independent units raise finite sample concerns for estimation and inference in Stage 2. Specifically, there can arise a tension between adjusting for too many partition-level covariates (with the potential of overfitting, even with cross-validation) and including too few (not supporting the identification assumptions). In future work, we plan to use "collaborative" TMLE (van der Laan and Gruber, 2010) where the partition-level propensity score would be fit in response to adjustment conducted in the partition-level outcome regression. As illustrated with the real-data example, in-depth discussion with subject matter experts is imperative to identifying the minimal adjustment set needed to support our assumptions — both on the missingness mechanism and on within cluster dependence. Conducting a simulation study, informed by the real-data application, can help guide development of the statistical analysis plan.

This work addresses four common challenges in the design and analysis of CRTs: (i) subsampling of participants for measurement of a rare or expensive outcome; (ii) missingness on the baseline outcome status of sampled participants; (iii) missingness on the final outcome status of participants known to be "at-risk" at baseline; and (iv) very few independent units (i.e., clusters). To address the first three challenges, we extended Two-Stage TMLE to account for potentially biased sampling and outcome measurement. To address the final challenge, we carefully articulated and critically evaluated the assumptions required to treat subcluster partitions as conditionally independent. These assumptions increase our effective sample size, at the cost of making the CRT behave more like an observational study.

ACKNOWLEDGEMENTS

We extend many thanks to the Ministries of Health of Uganda and Kenya; our research and administrative teams in San Francisco, Uganda, and Kenya; our collaborators and advisory boards; and, especially, all the communities and participants involved. We also thank Drs. Lina Montoya and Mark van der Laan for their expert advice. Funding support for this work was provided by The National Institutes of Health (U01AI099959, UM1AI068636, K23AI118592, and R01AI151209) and the President's Emergency Plan for AIDS Relief.

Funding

The National Institutes of Health (U01AI099959, UM1AI068636, K23AI118592, and R01AI 151209).

Supplementary Material

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

Conflict of Interest

None declared.

References

Andrews, J. R., Morrow, C., Walensky, R. P. and Wood, R. (2014). Integrating social contact and environmental data in evaluating tuberculosis transmission in a South African Township. *The Journal of Infectious Diseases* **210**, 597–603.

Balzer, L. B., Petersen, M. L., van der Laan, M. J. and the SEARCH Collaboration. (2016a). Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching. *Statistics in Medicine* **35**, 3717–3732.

Balzer, L. B., van der Laan, M., Ayieko, J., Kamya, M., Chamie, G., Schwab, J., Havlir, D. V. and Petersen, M. L. (2021). Two-Stage TMLE to reduce bias and improve efficiency in cluster randomized trials. *Biostatistics* **24**, 502–517.

Balzer, L. B., van der Laan, M. J. and Petersen, M. L. (2016b). Adaptive pre-specification in randomized trials with and without pair-matching. *Statistics in Medicine* **35**, 4528–4545.

Balzer, L. B., Zheng, W., van der Laan, M. J. and Petersen, M. L. (2019). A new approach to hierarchical data analysis: targeted maximum likelihood estimation for the causal effect of a cluster-level exposure. *Statistical Methods in Medical Research* **28**, 1761–1780.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–972.

Benitez, A., Petersen, M. L., van der Laan, M. J., Santos, N., Butrick, E., Walker, D., Ghosh, R., Otieno, P., Waiswa, P. and Balzer, L. B. (2023). Defining and estimating effects in cluster randomized trials: a methods comparison. *Statistics in Medicine* **42**, 3443–3466.

Campbell, M. J. and Walters, S. J. (2014). *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research*, 1st edition. Chichester, West Sussex: Wiley.

Chamie, G., Kato-Maeda, M., Emperador, D. M., Wandera, B., Mugagga, O., Crandall, J., Janes, M., Marquez, C., Kamya, M. R., Charlebois, E. D. *and others*. (2018). Spatial overlap links seemingly unconnected genotype-matched TB cases in rural Uganda. *PLoS One* **13**, e0192666.

Chamie, G., Wandera, B., Marquez, C., Kato-Maeda, M., Kamya, M. R., Havlir, D. V. and Charlebois, E. D. (2015). Identifying locations of recent TB transmission in rural Uganda: a multidisciplinary approach. *Tropical Medicine & International Health* **20**, 537–545.

Díaz, I. (2019). Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics* **21**, 353–358.

Donner, A. and Klar, N. (2010). *Design and Analysis of Cluster Randomization Trials in Health Research*, 1st edition. London: Wiley.

Eldridge, S. and Kerry, S. (2012). *A Practical Guide to Cluster Randomised Trials in Health Services Research*, 1st edition. Chichester: Wiley.

Fiero, M. H., Huang, S., Oren, E. and Bell, M. L. (2016). Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials* **17**, 72.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2012). *Applied Longitudinal Analysis*, 2nd edition. Wiley.

Getahun, H., Gunneberg, C., Granich, R. and Nunn, P. (2010). HIV infection—associated tuberculosis: the epidemiology and the response. *Clinical Infectious Diseases* **50**(Supplement_3), S201–S207.

Havlir, D. V., Balzer, L. B., Charlebois, E. D., Clark, T. D., Kwarisiima, D., Ayieko, J., Kabami, J., Sang, N., Liegler, T., Chamie, G. *and others*. (2019). HIV testing and treatment with the use of a community health approach in rural africa. *New England Journal of Medicine* **381**, 219–229.

Hayes, R. J. and Moulton, L. H. (2009). *Cluster Randomised Trials*, 1st edition. Boca Raton: Chapman and Hall/CRC.

Hernán, M. A., Lanoy, E., Costagliola, D. and Robins, J. M. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology* **98**, 237–242.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Houben, R. M. G. J. and Dodd, P. J. (2016). The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. *PLoS Medicine* **13**, e1002152.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., Bruckner, T. and Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* **21**, 467–474.

Kahan, B. C., Forbes, G., Ali, Y., Jairath, V., Bremner, S., Harhay, M. O., Hooper, R., Wright, N., Eldridge, S. M. and Leyrat, C. (2016). Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials* **17**, 438.

Kreft, I. G. G. (1998). *Introducing Multilevel Modeling*, 1 edition. London; Thousand Oaks, CalifA: SAGE Publications Ltd.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

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.

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

Lönnroth, K., Williams, B. G., Stadlin, S., Jaramillo, E. and Dye, C. (2008). Alcohol use as a risk factor for tuberculosis – a systematic review. *BMC Public Health* **8**, 289.

MacPherson, P., Moshabela, M., Martinson, N. and Pronyk, P. (2009). Mortality and loss to follow-up among HAART initiators in rural South Africa. *Transactions of The Royal Society of Tropical Medicine and Hygiene* **103**, 588–593.

Marquez, C., Atukunda, M., Nugent, J. R., Charlebois, E., Chamie, G., Mwangwa, F., Ssemmondo, E., Kironde, J., Kabami, J., Owaraganise, A., *and others*. (2022). Impact of a community-wide HIV test and treat intervention on population-level tuberculosis transmission in rural Uganda. 24th International AIDS Conference, Montreal, Canada.

Martinez, L., Lo, N. C., Cords, O., Hill, P. C., Khan, P., Hatherill, M., Mandalakas, A., Kay, A., Croda, J., Horsburgh, C. R., Zar, H. J. *and others*. (2019). Paediatric tuberculosis transmission outside the household: challenging historical paradigms to inform future public health strategies. *The Lancet Respiratory Medicine* **7**, 544–552.

Martinez, L., Shen, Y., Mupere, E., Kizza, A., Hill, P. C. and Whalen, C. C. (2017). Transmission of mycobacterium tuberculosis in households and the community: a systematic review and meta-analysis. *American Journal of Epidemiology* **185**, 1327–1339.

Moore, K. L. and van der Laan, M. J. (2009). Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *Statistics in Medicine* **28**, 39–64.

Murray, D. M., Pals, S. L., George, S. M., Kuzmichev, A., Lai, G. Y., Lee, J. A., Myles, R. L. and Nelson, S. M. (2018). Design and analysis of group-randomized trials in cancer: a review of current practices. *Preventive Medicine* **111**, 241–247.

Oakes, J. M. (2004). The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology (with discussion). *Social Science & Medicine* **58**, 1929–1952.

Pan, W. and Wall, M. M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine* **21**, 1429–1441.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge, U.K. Cambridge University Press.

Puryear, S. B., Fatch, R., Beesiga, B., Kekibiina, A., Lodi, S., Marson, K., Emenyonu, N. I., Muyindike, W. R., Kwarisiima, D., Hahn, J. A. *and others*. (2021). Higher levels of alcohol use are associated with latent tuberculosis infection in adults living with human immunodeficiency virus. *Clinical Infectious Diseases* **72**, 865–868.

Rehm, J., Samokhvalov, A. V., Neuman, M. G., Room, R., Parry, C., Lönnroth, K., Patra, J., Poznyak, V. and Popova, S. (2009). The association between alcohol use, alcohol use disorders and tuberculosis (TB). A systematic review. *BMC Public Health* **9**, 450.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512.

Robins, J., Orellana, L. and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine* **27**, 4678–4721.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

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

Selvaraj, S. and Prasad, V. (2013). Characteristics of cluster randomized trials: are they living up to the randomized trial? *JAMA Internal Medicine* **173**, 313–315.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association* **101**, 1398–1407.

Tsiatis, A. A., Davidian, M., Zhang, M. and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine* **27**, 4658–4677.

van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics* **8**, 1–39. Article 9.

van der Laan, M. J. and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics* **6**, 1–68. Article 17.

van der Laan, M. J., Petersen, M. and Zheng, W. (2013). Estimating the effect of a community-based intervention with two communities. *Journal of Causal Inference* **1**, 83–106.

van der Laan, M. J. and Petersen, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics* **3**, 1–52. Article 3.

van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* **6**.

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics. New York, NY, USA: Springer.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.