UC Berkeley

Title

A new approach to hierarchical data analysis: Targeted maximum likelihood estimation for the causal effect of a cluster-level exposure

Permalink

https://escholarship.org/uc/item/4qn0h393

Journal

ISSN

Authors

Balzer, Laura B
Zheng, Wenjing
van der Laan, Mark J
et al.

Publication Date

2019-06-01

DOI

Peer reviewed

A new approach to hierarchical data analysis: Targeted maximum likelihood estimation for the causal effect of a cluster-level exposure

Laura B Balzer[1], **Wenjing Zheng**[2], **Mark J van der Laan**[3], and **Maya L Petersen**[3] **on behalf of the SEARCH study**

[1]Department of Biostatistics & Epidemiology, School of Public Health & Health Sciences, University of Massachusetts, Amherst, MA, USA

[2]Netflix, Los Gatos, CA, USA

[3]Division of Biostatistics, School of Public Health, University of California, Berkeley, CA, USA

Abstract

We often seek to estimate the impact of an exposure naturally occurring or randomly assigned at the cluster-level. For example, the literature on neighborhood determinants of health continues to grow. Likewise, community randomized trials are applied to learn about real-world implementation, sustainability, and population effects of interventions with proven individual-level efficacy. In these settings, individual-level outcomes are correlated due to shared cluster-level factors, including the exposure, as well as social or biological interactions between individuals. To flexibly and efficiently estimate the effect of a cluster-level exposure, we present two targeted maximum likelihood estimators (TMLEs). The first TMLE is developed under a non-parametric causal model, which allows for arbitrary interactions between individuals within a cluster. These interactions include direct transmission of the outcome (i.e. contagion) and influence of one individual's covariates on another's outcome (i.e. covariate interference). The second TMLE is developed under a causal sub-model assuming the cluster-level and individual-specific covariates are sufficient to control for confounding. Simulations compare the alternative estimators and illustrate the potential gains from pairing individual-level risk factors and outcomes during estimation, while avoiding unwarranted assumptions. Our results suggest that estimation under the sub-model can result in bias and misleading inference in an observational setting. Incorporating working assumptions during estimation is more robust than assuming they hold in the underlying causal model. We illustrate our approach with an application to HIV prevention and treatment.

1   Introduction

In many studies, individuals are grouped into clusters, such as households, clinics, or communities, and the objective is to learn the impact of an exposure naturally occurring or randomly assigned at the cluster-level. In observational settings, for example, there is a growing body of literature dedicated to understanding neighborhood determinants of health. [1–3] Likewise, cluster (group) randomized trials are increasingly implemented to learn about large-scale implementation as well as the direct, indirect, and population-level effects of interventions with proven individual-level efficacy. [4] Examples of ongoing cluster randomized trials include the SEARCH study, testing a community-based strategy for HIV prevention and treatment [5]; the CBIM study, testing a school-based program to prevent gender violence [6]; and the SHINE study, testing a household-based strategy to reduce Staphylococcus aureus infection. [7] In both observational and trial settings, individual-level outcomes may be correlated due to shared cluster-level factors, including the exposure, and causal interactions between individuals within clusters. In this paper, we aim to make full use of a hierarchical data structure to flexibly and efficiently estimate the effect of the cluster-based exposure, while avoiding unwarranted causal and statistical assumptions.

There is an extensive literature on the definition and estimation of the impact of cluster-based exposures or interventions. [4,8] Two popular approaches are random (mixed) effects models and generalized estimating equations (GEE). [9,10] For reviews of these methods, we refer the reader to Gardiner et al. and Hubbard et al., among others. [11,12] In these approaches, the causal effect of interest is defined as the coefficient for exposure in the outcome regression. For estimation and inference, these algorithms harness the pairing of individual-level risk factors and outcomes, while accounting for the correlation of outcomes within clusters. More recently, augmented-GEE has been proposed to increase precision in cluster randomized trials. [13,14] A potential short-coming of these approaches is their reliance on parametric regression models to define and estimate causal effects. In particular, background knowledge is rarely sufficient to justify the parametric models employed. In observational settings, this can result in ill-defined causal effects, biased estimates, and misleading inference. [12] In cluster randomized trials, this approach can result in efficiency losses.

In this manuscript, we begin by presenting a structural causal model to represent a general hierarchical data generating process. [15–17] This causal model is non-parametric and accounts for dependence in individual-level outcomes that may be induced by shared cluster-level factors and by causal interactions between individuals. [3,18–22] Throughout we assume independence between clusters. The causal model can incorporate, but does not require, assumptions reflecting the exposure assignment to clusters (e.g. randomization). Through interventions on this causal model, we generate counterfactuals and define the causal effect of interest without relying on parametric models. This approach ensures that the causal

effect corresponds to the underlying scientific question and is agnostic to data generating process (e.g. the presence or absence of informative cluster sizes[23]).

If the observed data are aggregated to the cluster-level, then estimation of the corresponding statistical parameter can proceed analogously to non-hierarchical data structures. For example, we could apply matching algorithms,[24–26] parametric G-computation,[27–30] inverse probability of treatment weighting (IPTW) estimators,[31–36] or double robust approaches,[17,37–41] such as targeted maximum likelihood estimation (TMLE). This aggregated data approach is straightforward and naturally respects the experimental (independent) unit as the cluster. Furthermore, this approach avoids unwarranted assumptions on the distribution of latent terms or on the dependence structure within a cluster. However, this approach ignores the pairing of the individual-level risk factors with individual-level outcomes.

As an alternative to approaches based on aggregated data, we develop two targeted maximum likelihood estimators (TMLEs) that leverage the hierarchical data structure by preserving the pairing of individual-level covariates and outcomes.[17,41] TMLE is a general framework for the construction of double robust, semi-parametric, efficient, substitution estimators. As applied to causal effect estimation in a single time point setting, the algorithm begins with an initial estimator of the *outcome regression*: the conditional mean outcome, given the exposure and baseline covariates. TMLE updates this initial estimator by incorporating information in the known or estimated *propensity score*: the conditional probability of receiving the exposure, given the covariates. These updated estimates are then plugged into the parameter mapping. TMLE is a substitution estimator, which improves its stability. Through its updating procedures, TMLE satisfies the efficient score equation, while guaranteeing parameter estimates respect known bounds (contrary to a direct estimating equation approach). As a result, TMLE is double robust, yielding a consistent estimate if either the outcome regression or the propensity score is estimated consistently, and efficient, achieving the lowest possible variance if both the outcome regression and propensity score are estimated consistently at reasonable rates. Finally, TMLE naturally integrates machine learning, while maintaining the basis for formal statistical inference.

In this manuscript, we first propose incorporating the pairing of individual-level covariates and outcomes to improve initial estimation of the outcome regression in a cluster-level TMLE (Section 3). Then in Section 4, we consider assumptions commonly made when estimating effects in hierarchical settings. Specifically, we assume that an individual's outcome is generated as a common function of the cluster-level covariates, cluster-level exposure and individual-specific covariates, but is not directly affected by the covariates of other individuals within his/her cluster (i.e. no covariate interference[42]). We further assume that the cluster-level and individual-specific covariates are sufficient to control for confounding. For the resulting statistical parameter, we present a second TMLE for this distinct estimation problem.

We compare the two TMLEs theoretically (Section 5) and with finite sample simulations (Section 6). They differ in their efficiency and in how they incorporate individual-level data. In particular, the assumptions in the more restrictive sub-model result in a lower efficiency bound and thus a potentially more precise TMLE than that developed under the larger

model. However, if these assumptions do not hold, the TMLE developed under this sub-model may be subject to bias and misleading inference in an observational setting and to inefficiency in a trial setting. Since these assumptions are often made when estimating the effects of cluster-level exposures, our findings may have implications beyond the Targeted Learning framework.

To illustrate the concepts in this paper, we consider a community-based strategy for intensified HIV testing with immediate initiation of antiretroviral therapy (ART) for all HIV-infected individuals. The premise of this "Test-and-Treat" strategy is to improve clinical outcomes among HIV-infected individuals and dramatically reduce their probability to transmission to others.[43–48] Our objective is estimate the impact of this strategy as compared to the standard of care on cumulative HIV incidence: the proportion of baseline HIV-uninfected individuals who become HIV-infected by the end of follow-up. Within a community, individual outcomes are expected to be correlated due to both shared community-level factors and causal interactions between individuals. The desire to capture the direct, indirect, total, and overall effects of this nature are a common motivation for focusing on evaluation of cluster-level rather than individual-level interventions.[3,4,19–22]

2 General hierarchical causal model

We begin by specifying a structural causal model for the process that generated data on each cluster (the experimental unit).[15,16] Throughout, we focus on the simple scenario where a cluster is first sampled from some target population, and then individuals within a cluster are selected for participation. In the running example, a study community is randomly selected from the target population of communities, and then baseline HIV-uninfected individuals are randomly sampled from that community. The number of individuals selected in each cluster could be fixed or could vary. The latter case may arise if underlying cluster sizes differ and all eligible individuals in each cluster are selected. Throughout, clusters are indexed with $j = \{1, \ldots, J\}$, and individuals are indexed with $i = \{1, \ldots, N_j\}$.

After selection of the study units, covariates are measured. These baseline characteristics may affect, but are not themselves affected by the exposure. Some characteristics might be aggregates of individual-level covariates, while others may be cluster-level covariates with no clear individual-level counterpart. The baseline characteristics are divided into two mutually exclusive sets. For cluster j, let E_j denote the vector of environmental factors shared by all cluster members, and $\mathbf{W}_j$ the matrix of individual-level characteristics. In our example, E_j could include baseline HIV prevalence and community size, while individual-level covariates $\mathbf{W}_j$ might include baseline risk behaviors and demographic data, such as age, sex, and marital status. If there are p such individual-level covariates, then $\mathbf{W}_j$ would be an ($N_j \times p$) matrix and W_{ij} would be the ($1 \times p$) vector of baseline characteristics for subject i in cluster j. Throughout $W_{i\cdot}$ denotes the ith individual's covariates from a randomly selected (or unspecified) cluster from the target population.

Next the exposure A is assigned or naturally occurs in each cluster. In our example, A_j is an indicator that the Test-and-Treat strategy is implemented in community j. The exposure received by cluster j might be randomly assigned or might depend on the covariates (E_j, $\mathbf{W}_j$).

Finally, the outcome $\mathbf{Y}_j = (Y_{ij} : i = 1, \ldots, N_j)$ is measured on all selected individuals in cluster j. Throughout, Y_i denotes the ith individual's outcome from a randomly selected (or unspecified) cluster. In the example, Y_i is an indicator that individual i becomes HIV-infected by the end of follow-up.

Causal relationships between these variables are specified through a directed acyclic graph (Figure 1) or non-parametric structural equations[15,16]

$$
\begin{aligned}
E &= f_E(U_E) \\
\mathbf{W} &= f_{\mathbf{W}}(E, U_{\mathbf{W}}) \\
A &= f_A(E, \mathbf{W}, U_A) \\
\mathbf{Y} &= f_{\mathbf{Y}}(E, \mathbf{W}, A, U_{\mathbf{Y}})
\end{aligned}
\tag{1}
$$

where $U = (U_E, U_{\mathbf{W}}, U_A, U_{\mathbf{Y}})$ denotes the set of unmeasured variables. This model states that the value of each variable on the left-hand side of an equation may be causally determined by the variables on the right hand side of the equation, including unmeasured sources of random variation U. This causal model (equation (1)) contains data generating structures corresponding to both randomized trials and observational settings. See online Appendix A for further details.

This model accounts for many possible sources of dependence between individuals within a cluster. For example, individual-level variables (covariates and outcomes) will be correlated due to shared measured and unmeasured factors. The model also allows for contagion: when an individual's outcome Y_{ij} may be affected by another's outcome Y_{kj} within cluster j.[20] Covariate interference is also consistent with this model: an individual's outcome Y_{ij} may be affected by another's covariates W_{kj}.[42] No assumptions are made about the structure or correlation of the unmeasured factors ($U_{\mathbf{W}}, U_{\mathbf{Y}}$) between individuals within a cluster. Thus, this general causal model covers a wide range of "dependent happenings".[20] It does, however, assume causal independence between distinct clusters (communities).

2.1 Counterfactuals and the target causal effect

Counterfactual outcomes are defined through modifications to the data generating process described by the causal model (equation (1)).[16,17] Replacing the structural equation f_A with the constant a generates the counterfactual random variable $\mathbf{Y}(a)$. Under assumptions linking the structural causal model to the observed data (stated explicitly below), $\mathbf{Y}_j(a) = (Y_{ij}(a) : i = 1, \ldots, N_j)$ can be interpreted as the vector of individual-level outcomes that would be observed for cluster j under exposure level a. As before, $Y_i(a)$ denotes the ith individual's counterfactual outcome for a randomly selected (unspecified) cluster. In the example, $Y_i(1)$ represents the final HIV status for subject i were his/her community to receive the Test-and-Treat strategy, irrespective of whether or not the community in fact received the intervention. Likewise, $Y_i(0)$ represents the final HIV status for subject i were his/her community to continue with the standard of care.

Let the cluster-level counterfactual outcome be the (weighted) mean outcome for the N_j individuals sampled from cluster j

$$Y_j^c(a) \equiv \sum_{i=1}^{N_j} \alpha_{ij} Y_{ij}(a) \quad (2)$$

for some user-specified set of weights such that $\sum_{i=1}^{N_j} \alpha_{ij} = 1$ When the sample size N_j varies, a natural choice for the weights is the inverse cluster-specific sample size: $\alpha_{ij} = 1/N_j$. When the individual-level index i is informative (e.g. in a repeated measures setting), other choices of the weight vector α might be preferred. To simplify exposition for the remainder of the article, we assume the weight $\alpha_{ij} = 1/N_j$ and the cluster-level outcome is the empirical mean of the individual-level outcomes. In the running example, $Y^c(a)$ is the counterfactual proportion of baseline HIV-uninfected individuals who would seroconvert during follow-up if the community received exposure $A = a$. In other words, $Y^c(a)$ is the counterfactual cumulative HIV incidence under exposure level $A = a$.

We focus on causal parameters defined in terms of the *treatment-specific mean*, the expected counterfactual cluster-level outcome if all clusters in the target population received the exposure $A = a$: $\mathbb{E}[Y^c(a)]$. The difference or ratio of these treatment-specific means defines a causal effect. For example, the population average treatment effect is given by $\mathbb{E}[Y^c(1)] - \mathbb{E}[Y^c(0)]$. For the running example, this causal effect evaluates the difference in HIV incidence if all communities in our target population implemented the Test-and-Treat strategy versus if all communities continued with the standard of care. Alternatively, we could define our parameter of interest as the causal risk ratio: $\mathbb{P}(Y^c(1) = 1) \big/ \mathbb{P}(Y^c(0) = 1)$. For conditions and interpretation in terms of a pooled individual-level causal effect, see online Appendix B.

2.2 Observed data and statistical model

For a randomly sampled cluster, the observed data are the measured environmental covariates, the measured individual-level covariates, the exposure, and the vector of individual-level outcomes: $O = (E, \mathbf{W}, A, \mathbf{Y})$. We define the observed cluster-level outcome as the empirical mean of the individual-level outcomes: $Y_j^c \equiv \sum_{i=1}^{N_j} \alpha_{ij} Y_{ij}$ with our choice of weights $\alpha_{ij} = 1/N_j$. We assume that the observed data $O_j : j = 1, \dots, J$ are generated by sampling J independent times from some distribution compatible with the causal model. Thereby, the causal model (equation (1)) implies a statistical model, which describes the set of possible distributions of O and is denoted $\mathscr{M}^I$. In many cases, the causal model does not place any restrictions on the set of observed data distributions, and the resulting statistical model is non-parametric. In other cases, such as a randomized trial, knowledge about the exposure assignment mechanism implies a semi-parametric statistical model. We use subscript 0 to denote the true distributions. The true distribution of the observed data, denoted $\mathbb{P}_0$, is an element of the statistical model $\mathscr{M}^I$.

2.3 Identifiability of the cluster-level causal effect

To write the treatment-specific mean as a function of the observed data distribution, we make two additional assumptions, analogous to common identifiability assumptions for non-hierarchical causal effects.[27] First, we assume that all the common causes of the cluster-based exposure A and the vector of individual-level outcomes $\mathbf{Y}$ are captured by the measured covariates $(E, \mathbf{W})$ (Supplementary Figure S1). In other words, we assume there is no unmeasured confounding: $A \perp\!\!\!\perp \mathbf{Y}(a) \mid E, \mathbf{W}$. In the HIV example, this assumption would hold by design if the Test-and-Treat intervention were randomly allocated among communities. Otherwise, measuring a rich set of determinants of HIV infection will increase the plausibility of this assumption.

We also need the positivity assumption (a.k.a. experimental treatment assignment assumption), which ensures that there is sufficient variability in the exposure value within all possible confounder strata: $\mathbb{P}_0(A = a \mid E = e, \mathbf{W} = \mathbf{w}) > 0$ $a.e.$. Under these assumptions, we have the hierarchical analogue to the G-computation identifiability result[27]: $\mathbb{E}[\mathbf{Y}(a)] = \mathbb{E}_0[\mathbb{E}_0(Y \mid A = a, E, \mathbf{W})]$. This provides us with a general identifiability result for the causal effect of cluster-level exposure a on any cluster-level outcome Y^C, which is some real valued function of the outcome vector $\mathbf{Y}$:

$$\Psi^I(\mathbb{P}_0)(a) \equiv \mathbb{E}_0[\mathbb{E}_0(Y^c \mid A = a, E, \mathbf{W})] \quad (3)$$

We can interpret the resulting statistical estimand as the expected cluster-level outcome, given the exposure and covariates, averaged (standardized) with respect to the covariate distribution in the population.

The randomization and positivity assumptions thus allow us to identify parameters of $\mathbb{E}[Y^c(a)]$, such as the population average treatment effect: $\mathbb{E}_0[\mathbb{E}_0(Y^c \mid A = 1, E, \mathbf{W}) - \mathbb{E}_0(Y^c \mid A = 0, E, \mathbf{W})]$. Likewise, for a binary outcome, we can identify the causal risk ratio as $\mathbb{E}_0[\mathbb{E}_0(Y^c \mid A = 1, E, \mathbf{W})] \big/ \mathbb{E}_0[\mathbb{E}_0(Y^c \mid A = 0, E, \mathbf{W})]$.

3 Estimation under the general hierarchical causal model

In the previous section, we defined the statistical estimand as a mapping from the statistical model to the parameter space: $\Psi^I: \mathcal{M}^I \to R$. Under the above randomization and positivity assumptions, the target parameter $\Psi^I(\mathbb{P}_0)(a)$ corresponds to the treatment-specific mean $\mathbb{E}[Y^c(a)]$, which can be used to define both absolute and relative effects.[49]

In this section, we review a targeted maximum likelihood estimator (TMLE) of the statistical parameter $\Psi^I(\mathbb{P}_0)(a)$ (equation (3)) based on J i.i.d. observations O from $\mathbb{P}_0 \in \mathcal{M}^I$. The efficient influence curve and cluster-level TMLE presented in this section are direct analogs to the standard individual-level TMLE described and implemented elsewhere.[17,50,51] We then discuss several approaches for nuisance parameter estimation. Our contribution is to

consider candidate estimators making full use of the hierarchical data structure (i.e. the pairing of individual-level risk factors and outcomes) during initial estimation of the conditional mean outcome.

Before proceeding, we introduce some additional notation. Let us denote the marginal distribution of the baseline covariates as $Q_{E,\mathbf{W}} \equiv \mathbb{P}(E, \mathbf{W})$ and the conditional mean of the cluster-level outcome given the exposure and covariates as $\bar{Q}^c(A, E, \mathbf{W}) \equiv \mathbb{E}(Y^c \mid A, E, \mathbf{W})$. The statistical parameter can thus be written as $\mathbb{E}[\mathbb{E}(Y^c \mid a, E, \mathbf{W})] = \sum_{e,\mathbf{w}} \bar{Q}^c(a, e, \mathbf{w})\mathbb{P}(e, \mathbf{w})$, where the summation generalizes to the integral for continuous covariates. This clarifies that the statistical parameter only depends on the observed data distribution through $Q = (Q_{E,\mathbf{W}}, \bar{Q}^c)$. For the targeting step, we will also need to estimate the cluster-level propensity score, denoted as $g^c(a \mid E, \mathbf{W}) \equiv \mathbb{P}(A = a \mid E, \mathbf{W})$. Without loss of generality, we assume that the cluster-level outcome Y^c is bounded between zero and one.[52] In the running example, Y^c is cumulative HIV incidence and thus a proportion.

3.1 The cluster-level TMLE

The efficient influence curve of Ψ^I at $\mathbb{P}_0$ is given by

$$D^I(\mathbb{P}_0)(O) = \frac{\mathbb{I}(A = a)}{g_0^c(A \mid E, \mathbf{W})}(Y^c - \bar{Q}_0^c(A, E, \mathbf{W})) + \bar{Q}_0^c(a, E, \mathbf{W}) - \Psi^I(\mathbb{P}_0)(a) \quad (4)$$

where $\bar{Q}_0^c(A, E, \mathbf{W})$ denotes the true conditional mean of the cluster-level outcome and $g_0^c(A \mid E, \mathbf{W})$, denotes the true cluster-level propensity score. This is a direct analog of the efficient influuence curve for the G-computation identifiability result for a non-hierarchical data setting.[17,41] The first component is the weighted deviations between the cluster-level outcome and its expectation given the exposure and covariates; the weights are the inverse of the cluster-level propensity score. The second component is the deviation between the conditional mean outcome and its expectation over the covariate distribution. The efficient score equation can be generated as a score of a fluuctuation of the covariate distribution and the conditional distribution of the cluster-level outcome, given the exposure and covariates. This is used in formulation of the targeting step in TMLE.[17,41]

Specifically, suppose we have an initial estimator $\widehat{\bar{Q}}^c(A, E, \mathbf{W})$ of the expected cluster-level outcome $\bar{Q}_0^c(A, E, \mathbf{W})$. The TMLE algorithm updates this initial estimator with information contained in the known or estimated propensity score $\hat{g}^c(a \mid E, \mathbf{W})$. To do so, we minimize a pre-specified loss function along a least favorable (with respect to the statistical estimand) sub-model through $\widehat{\bar{Q}}^c(A, E, \mathbf{W})$. We choose the negative log-likelihood loss function

$$-\mathscr{L}^c(\bar{Q}^c)(O) = Y^c\log[\bar{Q}^c(A, E, \mathbf{W})] + (1 - Y^c)\log[1 - \bar{Q}^c(A, E, \mathbf{W})] \quad (5)$$

and the logistic sub-model with fluctuation parameter ϵ

$$logit[\widehat{\overline{Q}}^c(\epsilon)] = logit[\widehat{\overline{Q}}^c] + \epsilon\widehat{H}^c \quad (6)$$

Where $logit(\cdot) = log(\cdot / 1 - \cdot)$ and the "clever covariate" $\widehat{H}^c = \frac{\mathbb{I}(A = a)}{\hat{g}^c(a \mid E, \mathbf{W})}$. At zero

fluctuation, the initial estimator is returned: $\widehat{\overline{Q}}^c(\epsilon = 0) = \widehat{\overline{Q}}^c$. Furthermore, the score spans the relevant component of the efficient influence curve (equation (4)) at any distribution $\mathbb{P}$ in our model $\mathscr{M}^I$.

The parametric sub-model (equation (6)) is used to target the initial estimator $\widehat{\overline{Q}}^c(A, E, \mathbf{W})$ of outcome regression. The amount of fluctuation (i.e. the coefficient ϵ) is estimated with maximum likelihood. Specifically, we run logistic regression of the cluster-level outcome Y^c on the clever covariate $\widehat{H}^c$ and use the *logit* of the initial estimator $\widehat{\overline{Q}}^c(A, E, \mathbf{W})$ as offset. Plugging in the estimated fluctuation parameter $\hat{\epsilon}$ provides an updated fit to the outcome regression:

$$\widehat{\overline{Q}}^{c*} = \widehat{\overline{Q}}^c(\hat{\epsilon}) = logit^{-1}[logit(\widehat{\overline{Q}}^c) + \hat{\epsilon}\widehat{H}^c] \quad (7)$$

As estimator of the covariate distribution, we use the empirical $\hat{Q}_{E,\mathbf{W}}$, which puts weight $1/J$ on each cluster. As detailed in Rose and van der Laan,[53] the empirical distribution solves the relevant score equation (i.e. relevant component of the efficient influence curve) and does not need targeting, even in high dimensional settings.[54]

The TMLE is the substitution estimator obtained by plugging $\hat{Q}^* = (\hat{Q}_{E,\mathbf{W}}, \widehat{\overline{Q}}^{c*})$, into the parameter mapping Ψ^I:

$$\Psi^I(\hat{Q}^*)(a) = \frac{1}{J}\sum_{j=1}^{J}\widehat{\overline{Q}}^{c*}(a, E_j, \mathbf{W}_j) \quad (8)$$

The point estimate, denoted $\hat{\psi}^I(a)$, is the sample average of the targeted predictions of the cluster-level outcome, given the exposure of interest ($A = a$) and the measured covariates.

By construction, TMLE solves the efficient score equation: $0 = \sum_j D^I(\hat{Q}^*, \hat{g}^c)$. As a result, the estimator is double robust in that it remains consistent if only one of the nuisance parameters (the outcome regression or the propensity score) is consistently estimated. In an observational setting, this double robustness property improves our chances for obtaining a consistent estimate and valid statistical inference.[55] In a randomized trial, where the propensity score is known, the double robustness property implies that the TMLE will remain unbiased regardless of the outcome regression specification and thereby confers

wider flexibility in covariate adjustment to increase efficiency.[56] Furthermore, if both nuisance parameters are consistently estimated at reasonable rates,[17] then the TMLE is asymptotically linear with influence curve equal to the efficient influence curve (equation (4)) and asymptotically efficient.[57] In other words, this TMLE achieves the lowest possible asymptotic variance among a large class of estimators.

Under more general conditions,[17] TMLE is a regular, asymptotically linear estimator, and the Central Limit Theorem can be used to obtain statistical inference. Specifically, let

$$\hat{D}^I(\hat{Q}^*, \hat{g}^c)(O_j) = \frac{\mathbb{I}(A_j = a)}{\hat{g}^c(A_j \mid E_j, \mathbf{W}_j)}(Y_j^c - \widehat{\bar{Q}}^{c\,*}(A_j, E_j, \mathbf{W}_j)) + \widehat{\bar{Q}}^{c\,*}(a, E_j, \mathbf{W}_j) - \hat{\psi}^I(a) \quad (9)$$

be the plug-in estimator of the influence curve for observation O_j. We obtain a variance estimator with the sample variance of $\hat{D}^I(\hat{Q}^*, \hat{g}^c)$) divided by the number of experimental units: $\hat{\sigma}^2 = Var[\hat{D}^I] \big/ J$. This variance estimator is used to construct Wald-Type 95%-confidence intervals and carry out hypothesis tests. Under additional assumptions, the non-parametric bootstrap provides an alternative to the influence curve-based inference.

3.2   Data-adaptive estimation of nuisance parameters

In most applied settings, *a priori*-specification of a correct parametric regression for the conditional mean outcome $\bar{Q}_0^c(A, E, \mathbf{W})$ is impossible. We may know and measure the relevant covariates, but specifying the exact functional form is beyond our knowledge. (Recall our causal model often implies a non-parametric or semi-parametric statistical model.) In a randomized trial, the propensity score is known (e.g. $g_0^c(a \mid E, \mathbf{W}) = 0.5$)) and can be consistently estimated with a parametric regression to improve precision.[36,58,59] In observational settings, however, consistent estimation of the propensity score may present similar challenges. A core feature of TMLE is the use of machine learning algorithms for estimation of both the outcome regression $\bar{Q}_0^c(a, E, \mathbf{W})$ and the propensity score $g_0^c(a \mid E, \mathbf{W})$.

We focus on Super Learner,[60,61] an ensemble algorithm.[62,63] Super Learner employs V-fold cross-validation to build a convex combination of algorithm-specific predictions to minimize the cross-validated risk, based on a user-specified loss function. The library of candidate algorithms can include both parametric models and data-adaptive methods (e.g. stepwise regression, support vector machines,[64] generalized additive models,[65] LASSO[66] – each with multiple tuning parameters). If the correctly specified parametric model is not included in the library, Super Learner under minimal conditions performs asymptotically as well as the "oracle selector" that uses the true distribution $\mathbb{P}_0$ to select the optimal convex combination from the library.[60,61] If the correctly specified parametric model is included in the library, Super Learner still achieves an almost parametric rate of convergence.

Under our statistical model $\mathcal{M}^I$, Super Learner for the outcome regression and for the propensity score can be implemented using a cluster-level loss function (online Appendix

C). Alternatively, to leverage the pairing of individual-level covariates and outcomes and to reduce the dimensionality of the adjustment set, we now consider two working assumptions. These assumptions suggest alternative approaches to estimating the cluster-level outcome regression $\bar{Q}_0^c(A, E, \mathbf{W})$ and thereby an expanded Super Learner library.

First, suppose that an individual's outcome is minimally impacted by the covariates of other individuals in his or her cluster: $\mathbb{E}_0(Y_{i\cdot} \mid A, E, \mathbf{W}) = \mathbb{E}_0(Y_{i\cdot} \mid A, E, W_{i\cdot})$. In other words, consider an exclusion restriction that the ith individual's outcome $Y_{i\cdot}$ is only a function of the matrix $\mathbf{W}$ through his/her own covariates $W_{i\cdot}$. Second, suppose that this individual-level regression is common in i: $\mathbb{E}_0(Y_{i\cdot} \mid A, E, W_{i\cdot}) = \bar{Q}_0(A, E, W)$ for some function $\bar{Q}_0$. A common function is natural when i indexes a random permutation $\{1, \ldots, N\}$. Under these working assumptions, we can rewrite the conditional mean of the cluster-level outcome as

$$\bar{Q}_0^c(A, E, \mathbf{W}) = \sum_{i=1}^{N} \alpha_{i\cdot} \bar{Q}_0(A, E, W_{i\cdot}) \quad (10)$$

This suggests a natural estimator for $\bar{Q}_0^c$ based on fitting a single regression of the individual-level outcome Y on the exposure and covariates (A, E, W) and then averaging across individuals within a cluster. In our HIV example, we could estimate the expected cumulative HIV incidence $\bar{Q}_0^c$ by (i) pooling individuals across clusters, (ii) fitting a individual-level outcome regression with weights a_{ij} and with terms for the cluster-level exposure, the community's baseline HIV prevalence as well as the individual's age and sex; and (iii) averaging the individual-level predictions within clusters. Corresponding data-adaptive approaches are also possible.

These working assumptions can be relaxed by incorporating knowledge of the dependence structure between individuals within clusters. Suppose we are able to identify or approximate for each individual i the specific set of individuals $C_{i\cdot}$ to which individual i is "connected". In other words, $C_{i\cdot}$ denotes the subset of individuals who influence the ith individual's outcome $Y_{i\cdot}$. Then we could pose a more general version of the working model (equation (10)) by including in the ith individual's covariate vector the covariates of his/her connections $W_{k\cdot}$ for $k \in C_{i\cdot}$. In the HIV example, an individual's probability of seroconversion might depend on his/his own sexual behavior as well as the baseline behavior of the other individuals in his/her sexual network $W_{k\cdot} : k \in C_{i\cdot}$.

In summary, the utility of the working assumption (equation (10)) is to generate an expanded set of candidate estimators of the conditional mean of the cluster-level outcome $\bar{Q}_0^c(A, E, \mathbf{W})$ for inclusion in the Super Learner library. Any $(N \times 1)$ individual-level covariate vector can alternatively be included in either the covariate matrix $\mathbf{W}$ or as a cluster-level covariate E. Therefore, we can include algorithms that assume $Y_{i\cdot}$ only depends on $W_{i\cdot}$ for investigator-specified subsets of $\mathbf{W}$. In other words, this working model allows us to consider a variety of dimension reductions for the adjustment set $(E, \mathbf{W})$. Super Learner provides a mechanism to

choose between and combine candidate individual-level and cluster-level algorithms in response to the data, thereby optimizing estimator performance. Step-by-step implementation of the cluster-level TMLE with Super Learner and corresponding *R* code is given in online Appendix D.

4   Hierarchical TMLE when causal dependence is restricted

The cluster-level TMLE, presented in the previous section, is developed under a general hierarchical causal model that makes no assumptions about the nature or sources of dependence between individuals within a cluster (equation (1); Figure 1). For identifiability of the impact of the cluster-level exposure, we assume the cluster-level covariates E and whole matrix of individual-level covariates $\mathbf{W}$ are sufficient to control for confounding. For initial estimation of the conditional mean of the cluster-level outcome $\bar{Q}_0^c(A, E, \mathbf{W})$, we consider additional working assumptions designed to more fully leverage the hierarchical nature of the data (equation (10)). These assumptions are treated as "working" and are not considered to reflect the underlying causal process. If the propensity score $g_0^c(A \mid E, \mathbf{W})$ is estimated consistently (as will always be true in a randomized trial), then estimating the outcome regression $\bar{Q}_0^c(A, E, \mathbf{W})$ under these working assumptions may improve asymptotic efficiency as well as finite sample bias and variance; the better the working assumptions approximate the truth, the better the TMLE will perform.

In this section, we consider an alternative hierarchical causal model, which restricts the causal dependence of individuals within a cluster. Specifically, we assume that an individual's outcome is known not to be affected by the covariates of other individuals in the cluster. This more restrictive causal model implies that the working assumptions (equation (10)) hold, thereby changing the statistical model by restricting the set of allowed distributions for outcome regression $\bar{Q}_0^c$. The modified causal model also results in a distinct identifiability result and corresponding statistical estimand. Specifically, we now need to assume that the cluster-level covariates E and individual *i*-specific covariates $W_{i.}$ are sufficient to control for confounding. For the modified statistical estimation problem, we present the efficient influence curve and the corresponding individual-level TMLE.

4.1   Restricted hierarchical causal model

We now consider a causal model assuming each individual's outcome $Y_{i.}$ is drawn from a common (in *i*) distribution depending on the cluster-level covariates E, each individual's own covariates $W_{i.}$, the cluster-level exposure A, and unmeasured factors $U_{Yi.}$, but not on the measured covariates of all other individuals in that cluster. In other words, we assume no covariate interference.[42] We further assume that the cluster-level covariates E and individual *i*-specific covariates $W_{i.}$ are sufficient to control for confounding. This assumption holds by design in a randomized trial, but is a strong assumption on the distribution of unmeasured factors in an observational setting (Supplementary Figure S2).

This data generating process is represented by the following structural causal model

$$E = f_E(U_E)$$
$$\mathbf{W} = f_\mathbf{W}(E, U_\mathbf{W})$$
$$A = f_A(E, \mathbf{W}, U_A) \qquad (11)$$
$$Y_{i.} = f_Y(E, W_{i.}, A, U_{Y_{i.}}), i = 1, \ldots, N$$

$$\text{where} A \perp\!\!\!\perp Y_{i.}(a) \mid E, W_{i.}$$

We further assume that the conditional probability distributions of the individual-level covariates and outcome (W_i, Y_i), given the cluster-level covariates and exposure (E, A), are common in i. This causal model is compatible with observational studies (Figure 2) and cluster randomized trials (Supplementary Figure S3).

Returning to our HIV example, the restricted causal model (equation (11)) assumes individual i's final HIV status Y_i is generated as a common function of the shared environmental factors E (e.g. region, baseline prevalence), his/her own covariates W_i (e.g. age, sex, marital status), implementation of the Test-and-Treat strategy A, and unmeasured individual-level factors U_{Y_i} (e.g. his/her perceived stigma), but not the covariates of others in his/her cluster. In this infectious disease setting, the restricted causal model might not be realistic. First, the baseline risk behavior of one individual W_k may directly or indirectly impact the outcome of another Y_i. Even if the assumption of no covariate interference is plausible, this causal model will not hold if there is an unmeasured common cause (e.g. community-level stigma) of the individual-level covariates (W_i, W_k) and outcomes (Y_i, Y_k). Of course, we could improve plausibility of these assumptions by including in W_i the baseline covariates of his/her partners C_i. Nonetheless, the assumptions in the restricted causal model are commonly made, but potentially implausible when outcomes are biologically or socially transmitted. We refer the reader to Supplementary Figure S2 for additional examples and discussion.

4.2 The individual-level TMLE

Assuming the restricted causal model is true, we proceed to estimation. As before, the observed data consist of J i.i.d observations of $O = (E, \mathbf{W}, A, \mathbf{Y})$, and the observed cluster-level outcome is the empirical mean of the individual-level outcomes: $Y^c = \sum_{i=1}^{N} \alpha_i \cdot Y_i$.

Without loss of generality, we assume that the individual-level outcome Y_i is bounded in zero and one.[52] In the example, Y_i is an indicator that the ith individual becomes infected with HIV over the course of follow-up.

The restricted causal model (equation (11)) implies the statistical assumption in equation (10); the conditional mean of the cluster-level outcome can be written as the average of individual-level regressions. We further assume that the conditional distribution of the exposure, given the cluster-level and individual i-specific covariates, is a common conditional distribution:

$$\mathbb{P}_0(A \mid E, W_{i\cdot}) \equiv g_0(A \mid E, W_{i\cdot}) \quad (12)$$

We refer to $g_0(A|E, W_{i\cdot})$ as the individual-level propensity score. The resulting statistical model implied by these assumptions is denoted $\mathcal{M}^{II}$ and is a sub-model of $\mathcal{M}^I$.

Under this more restrictive causal model, adjustment for the cluster-level covariates E and the individual i-specific covariates $W_{i\cdot}$ is sufficient to control for confounding. With the corresponding positivity assumption, our identifiability result for the treatment-specific mean $\mathbb{E}[Y^c(a)]$ is given by by

$$\Psi^{II}(\mathbb{P}_0)(a) \equiv \mathbb{E}_0\left\{ \sum_{i=1}^{N} \alpha_{i\cdot} \bar{Q}_0(a, E, W_{i\cdot}) \right\} \quad (13)$$

Let $\Psi^{II}: \mathcal{M}^{II} \to \mathbb{R}$ be the statistical parameter implied by this identifiability result, thus defining a new statistical estimation problem. As before, the statistical estimand $\Psi^{II}(\mathbb{P})(a)$ depends on the observed data distribution $\mathbb{P}$ through the marginal distribution of baseline covariates and the conditional mean of the cluster-level outcome: $Q = (Q_{E,\mathbf{W}}, \bar{Q}^c)$. Now, however, the conditional mean of the cluster-level outcome is assumed to be an average of common individual-level regressions: $\bar{Q}_0^c(A, E, \mathbf{W}) = \sum_i \alpha_{i\cdot} \bar{Q}_0(A, E, W_{i\cdot})$.

The efficient influence curve of ψ^{II} at $\mathbb{P}_0 \in \mathcal{M}^{II}$ is given by

$$D^{II}(\mathbb{P}_0)(O) = \sum_{i=1}^{N} \alpha_{i\cdot} \left(\frac{\mathbb{I}(A=a)}{g_0(A \mid E, W_{i\cdot})} (Y_{i\cdot} - \bar{Q}_0(A, E, W_{i\cdot})) + \bar{Q}_0(a, E, W_{i\cdot}) - \Psi^{II}(\mathbb{P}_0)(a) \right)$$

$$(14)$$

Under sub-model $\mathcal{M}^{II}$, the efficient influence curve is the average of an individual-level function. The first component of this individual-level function is the weighted deviations between the individual-level outcome and its expectation given the exposure and covariates; the weight is the inverse of the individual-level propensity score. The second component is the deviation between the conditional expectation of the individual-level outcome and the target parameter.

As before, the efficient influence curve (equation (14)) is used to derive the TMLE. Specifically, suppose we have an initial estimator $\widehat{\bar{Q}}(A, E, W)$ of the individual-level outcome regression $\bar{Q}_0(A, E, W)$ and an estimator $\hat{g}(a \mid E, W)$ of the individual-level propensity score $g_0(a|E, W)$. The TMLE algorithm updates the initial estimator $\widehat{\bar{Q}}(A, E, W)$ into $\widehat{\bar{Q}}^*(A, E, W)$ by

minimizing a pre-specified loss function along a least favorable (with respect to the statistical estimand) sub-model through $\widehat{\bar{Q}}(A, E, W)$. This updating step also serves to target the initial cluster-level outcome regression $\widehat{\bar{Q}}^c = \sum_i \alpha_i . \widehat{\bar{Q}}$ into $\widehat{\bar{Q}}^{c*} = \sum_i \alpha_i . \widehat{\bar{Q}}^*$.

As loss function for the outcome regression, we use the average of an i-specific loss function

$$\mathscr{L}^{II}(\bar{Q}^c)(O) = \sum_{i=1}^{N} \alpha_i . \mathscr{L}(\bar{Q})(O) \quad (15)$$

where

$$-\mathscr{L}(\bar{Q})(O) = Y_{i.} \log[\bar{Q}(A, E, W_{i.})] + (1 - Y_{i.}) \log[1 - \bar{Q}(A, E, W_{i.})] \quad (16)$$

$\mathscr{L}$ is a valid loss function for the i-specific outcome regression $\mathbb{E}_0(Y_{i.} \mid A, E, W_{i.})$, and under the sub-model $\mathscr{M}^{II}$ this regression is constant across individuals $\bar{Q}_0(A, E, W)$. Therefore, this is a valid loss function for each i, and the sum loss is also valid (online Appendix C).

For our fluctuation model through an initial estimator $\widehat{\bar{Q}}(A, E, W)$, we select the individual-level analog to the cluster-level fluctuation model (equation (6))

$$logit[\widehat{\bar{Q}}(\epsilon)] = logit[\widehat{\bar{Q}}] + \epsilon \widehat{H} \quad (17)$$

where the individual-level clever covariate is defined as

$$\widehat{H}_{ij} = \frac{\mathbb{I}(A_j = a)}{\widehat{g}(a \mid E_j, W_{ij})}, i = 1, \dots, N_j \text{ and } j = 1, \dots, J \quad (18)$$

This fluctuation model is only a function of the covariate matrix $\mathbf{W}$ through the ith-specific covariate $W_{i.}$ and is a sub-model of $\mathscr{M}^{II}$. At zero fluctuation, the initial estimator is returned. This combination of loss function and fluctuation model has score $\frac{d}{d\epsilon}\mathscr{L}^{II}(\bar{Q}^c)(\epsilon)$ at $\epsilon = 0$ that spans the relevant portion of the efficient influence curve D^{II}.

The amount of fluctuation ϵ is fit by pooling individuals across clusters and running logistic regression of the individual-level outcome $Y_{i.}$ on the clever covariate $\widehat{H}_{i.}$ with the *logit* of the initial estimator $\widehat{\bar{Q}}(A, E, W_{i.})$ as offset and weights $\alpha_{i.}$. Plugging in the resulting coefficient estimate $\hat{\epsilon}$ provides an updated fit of the individual-level regression

$$\widehat{\overline{Q}}^* = \widehat{\overline{Q}}(\hat{\epsilon}) = logit^{-1}[logit(\widehat{\overline{Q}}) + \hat{\epsilon}\widehat{H}] \quad (19)$$

and thereby the cluster-level regression: $\widehat{\overline{Q}}^{c\,*}(A, E, \mathbf{W}) = \sum_{i=1}^{N} \alpha_i \widehat{\overline{Q}}^*(A, E, W_{i\cdot})$.

As an initial estimator of the covariate distribution, we again use the empirical distribution $\widehat{Q}_{E,\mathbf{W}}$, which puts weight $1/J$ on each cluster. As before, the empirical distribution $\widehat{Q}_{E,\mathbf{W}}$ is the non-parametric maximum likelihood estimator and does not need to be targeted.[53,54] Therefore, the TMLE is defined as the substitution estimator obtained by plugging $\widehat{Q}^* = (\widehat{Q}_{E,\mathbf{W}}, \widehat{\overline{Q}}^{c\,*})$ into the parameter mapping Ψ^{II}

$$\Psi^{II}(\widehat{Q}^*)(a) = \frac{1}{J}\sum_{j=1}^{J}\left\{\sum_{i=1}^{N_j}\alpha_{ij}\widehat{\overline{Q}}^*(a, E_j, W_{ij})\right\} = \frac{1}{J}\sum_{j=1}^{J}\widehat{\overline{Q}}^{c\,*}(a, E_j, \mathbf{W}_j) \quad (20)$$

The point estimate, denoted $\hat{\psi}^{II}(a)$, is the sample average of the targeted predictions of the cluster-level outcome, given the exposure of interest ($A=a$) and the measured covariates. By construction, this TMLE solves the efficient influence curve equation: $\sum_j D^{II}(\widehat{Q}^*, \hat{g})(O_j) = 0$. Thereby, the estimator is double robust and asymptotically efficient under consistent estimation of both the outcome regression and propensity score.

Statistical inference proceeds as presented in Section 3. Specifically, let

$$\widehat{D}^{II}(\widehat{Q}^*, \hat{g})(O_j) = \sum_{i=1}^{N_j}\alpha_{ij}\left(\frac{\mathbb{I}(A_j = a)}{\hat{g}(A_j \mid E_j, W_{ij})}(Y_{ij} - \widehat{\overline{Q}}^*(A_j, E_j, W_{ij})) + \widehat{\overline{Q}}^*(a, E_j, W_{ij}) - \hat{\psi}^{II}(a)\right)$$

$$(21)$$

be the plug-in estimator of the influence curve for observation O_j. We obtain a variance estimator with the sample variance of $\widehat{D}^{II}(\widehat{Q}^*, \hat{g})$ divided by the number of experimental units: $\hat{\sigma}^2 = Var[\widehat{D}^{II}]\big/ J$. For this sub-model, an alternative variance estimator, which explicitly estimates the correlation structure within each cluster, is proposed in Schnitzer et al.[67]

Online Appendix D provides step-by-step implementation of the individual-level TMLE and corresponding R code. This individual-level TMLE can also be implemented with the existing ltmle[51] package using id to specify the clusters (independent units) and observation.weights for the weights α_{ij}. It is worth emphasizing, however, that this individual-level TMLE for the impact of a cluster-level exposure is developed under a causal

model with strong assumptions (equation (11)). In the following sections, we explore the theoretical and practical consequences of these assumptions.

5 Theoretical comparison of the TMLEs

The cluster-level TMLE is derived under a general causal model allowing for arbitrary dependence of individuals within a cluster. Our contribution is to propose incorporating pooled individual-level regressions as candidates in the Super Learner library for initial estimation of the expected cluster-level outcome $\bar{Q}_0^c(a, E, \mathbf{W})$. In contrast, the individual-level TMLE is derived under the restricted causal model, which assumes that the covariates of one individual do not affect the outcome of another (i.e. no covariate interference) and that the cluster-level covariates E and individual i-specific covariates W_i are sufficient to control for confounding. In practice, implementation of the two estimators differs in where and when we take averages. In the larger model, we immediately average any individual-level regressions to obtain an initial estimator $\widehat{\bar{Q}}^c(a, E, \mathbf{W})$ and target using a cluster-level clever covariate. In the sub-model, we update the individual-level estimator $\widehat{\bar{Q}}(a, E, W)$ using an individual-level clever covariate and then average the targeted predictions within each cluster: $\widehat{\bar{Q}}^{c*}(a, E, \mathbf{W}) = \sum_i \alpha_i . \widehat{\bar{Q}}^*(a, E, W_i.)$.

To compare the asymptotic efficiency of the two approaches, we first consider the special case where the exposure assignment A is independent of the whole covariate matrix $\mathbf{W}$, given the environmental factors E. In the running example, this would hold by design if the Test-and-Treat intervention were randomized $g_0^c(A \mid E, \mathbf{W}) = 0.5$. More generally, this condition would hold if the intervention were rolled out according only to community-level characteristics, such as baseline HIV prevalence and perceived need: $g_0^c(A \mid E, \mathbf{W}) = g_0^c(A \mid E)$. In this case, the efficiency bound for $\Psi^I(\mathbb{P}_0)$, presented in Section 3, will be identical to the efficiency bound for $\Psi^{II}(\mathbb{P}_0)$, presented in Section 4. In other words, we have the efficient influence curves are equal: $D^I(\mathbb{P}_0) = D^{II}(\mathbb{P}_0)$ at a $\mathbb{P}_0 \in \mathcal{M}^{II}$ (Proof in online Appendix E).

However, this does not imply that the corresponding TMLEs will be identical if the propensity score is unknown. In an observational setting, estimating a cluster-level propensity score $g_0^c(a \mid E, \mathbf{W})$ when implementing the TMLE for $\Psi^I(\mathbb{P}_0)$ as compared to an individual-level propensity score $g_0(a|E, W)$ when implementing the TMLE for $\Psi^{II}(\mathbb{P}_0)$ can result in estimators that are asymptotically distinct. If on the other hand, the exposure mechanism depends on both the environmental factors and the covariate matrix $g_0^c(A, E, \mathbf{W}) \neq g_0^c(A \mid E)$, then the efficiency bound for $\Psi^{II}(\mathbb{P}_0)$ in the smaller model $\mathcal{M}^{II}$ will be better than the efficiency bound for $\Psi^I(\mathbb{P}_0)$ in the larger model $\mathcal{M}^I$.

6 Finite sample simulations

In this section, we investigate the practical performance of the two TMLEs. We begin with a simple simulation to demonstrate implementation and performance in an observational setting. We then present a more realistic simulation, generated to reflect the HIV prevention and treatment example. Throughout, the causal parameter is the population average treatment effect $\mathbb{E}[Y^c(1) - Y^c(0)]$. All simulations were conducted using R.[68] Full computing code is publicly available.

6.1 Simulation 1 – Simple observational setting

We consider a sample size of $J = 100$ clusters. For each unit $j = \{1, \ldots, J\}$, we draw the number of individuals N_j from a normal with mean 50 and standard deviation 10 and round to the nearest whole number. Then for each individual $i = \{1, \ldots, N_j\}$, two covariates ($W1$, $W2$) are drawn from a multivariate normal. We include their averages as cluster-level covariates: $W1_j^c = 1 \left/ N_j \sum_i W1_{ij} \right.$ and $W2_j^c = 1 \left/ N_j \sum_i W2_{ij} \right.$. We consider an observational setting where the propensity score depends on one cluster-level aggregate: $A_j \sim Bern(logit^{-1}(0.75 W1_j^c))$. The probability of the individual-level outcome is simulated under two data generating distributions. Specifically, we vary the strength of the coefficients to simulate scenarios with minimal covariate interference

$$\mathbb{P}_0(Y_{ij} = 1 \mid A_j, W1_j^c, W2_j^c, W1_{ij}, W2_{ij}) = logit^{-1}(0.25 + 0.1 A_j + 0.15 W1_j^c + 1.15 W1_{ij} + W2_{ij}) \quad (22)$$

and with stronger covariate interference

$$\mathbb{P}_0(Y_{ij} = 1 \mid A_j, W1_j^c, W2_j^c, W1_{ij}, W2_{ij}) = logit^{-1}(0.25 + 0.1 A_j + 0.15 W1_j^c + 0.25 W1_{ij} + W2_j^c)$$

$$(23)$$

In the first data generating process (equation (22)), an individual's outcome is strongly impacted by his/her own covariates ($W1_{ij}$, $W2_{ij}$) and only weakly impacted by the covariates of others (i.e. $W1_j^c$). In the second process (equation (23)), the opposite holds. We then simulate the binary individual-level outcome as

$$Y_{ij} = \mathbb{I}(U_{Y_{ij}} < \mathbb{P}_0(Y_{ij} \mid A_j, W1_j^c, W2_j^c, W1_{ij}, W2_{ij})) \quad (24)$$

where the unmeasured error $U_{Y_{ij}} \in [0,1]$ is generated under two scenarios: independent within a cluster and correlated within a cluster. In the former, $U_{\mathbf{Y}_j} = (U_{Y_{ij}} : i, \ldots, N_j)$ is

generated by independently drawing N_j times from a Uniform(0,1), while in the latter U_{Y_j} is generated by applying the cumulative distribution function to correlated normal random variables (Full R code in Appendix D). Varying the dependence of the unmeasured factors U_Y determining the outcomes Y within a cluster allows us to examine the randomization assumption inherent in the restricted causal model (equation (11)). In practice, independent U_Y might be reasonable for outcomes that are not biologically or socially transmitted, but may be unreasonable otherwise (Supplementary Figure S2).

As before, we define the cluster-specific outcome Y^c as the empirical mean of the individual-level outcomes within that cluster. We generate counterfactual outcomes ($Y^c(1)$, $Y^c(0)$) by setting the cluster-level exposure to $A = 1$ and $A = 0$, respectively. For each data generating process, the average treatment effect is calculated by taking the mean difference in the counterfactual cluster-level outcomes for a population of 10,000 clusters (Supplementary Table S1). We also simulate under the null by setting the counterfactual outcome under the intervention equal to counterfactual outcome under the control.

As shown in Table 1, we consider three targeted estimators: TMLE-*Ia* adjusting for the covariates at the cluster-level in both the outcome regression and the propensity score regression; TMLE-*Ib* adjusting at the individual-level in the outcome regression and at the cluster-level in the propensity score regression; and TMLE-*II* adjusting at the individual-level in both the outcome regression and propensity score regression. TMLE-*Ia* and TMLE-*Ib* correspond to statistical model $\mathcal{M}^I$ and TMLE-*II* to sub-model $\mathcal{M}^{II}$. Both TMLE-*Ib* and TMLE-*II* harness the pairing of individual-level covariates and outcomes, but the former incorporates this information as working assumptions (equation (10)) during the estimation step, while the latter assumes the restricted causal model (equation (11)) reflects the true data generating process. We compare the targeted estimators to the unadjusted estimator, the average difference in cluster-level outcomes between treated and control groups.

6.1.1 Results—Table 2 provides a summary of the estimator performance over 5000 repetitions of the simulation. Recall the unadjusted estimator is simple the difference in average outcomes among treated units and average outcomes among control units. TMLE-*Ia*, developed under the general model $\mathcal{M}^I$, corresponds to an aggregated data approach; cluster-level regressions are used for both initial estimation and targeting. TMLE-*Ib*, also developed under the general model $\mathcal{M}^I$, uses a pooled individual-level regression for initial estimation of the mean outcome and then a cluster-level regression for updating. TMLE-*II*, developed under the more restrictive sub-model $\mathcal{M}^{II}$, uses pooled individual-level regressions for both initial estimation and updating.

When the unmeasured factors determining the outcome U_Y are independent, the unadjusted estimator, which fails to control for measured confounding, is biased. This bias is substantial enough to prevent reliable inference; the 95% confidence interval coverage is < 50%. The TMLE corresponding to an aggregated data approach (TMLE-*Ia*) performs well with negligible bias and good confidence interval coverage. When there is minimal covariate interference (equation (22)), TMLE-*Ib*, which makes working assumptions for initial estimation of the outcome regression, performs similarly to TMLE-*Ia*. However, when there

is stronger covariate interference (equation (23)) and these working assumptions fail, TMLE-*Ib* provides less power (23% vs. 34%) and conservative confidence interval coverage (98%). In this scenario, the cluster-level regression provides a better approximation of the true outcome regression resulting in greater efficiency and power for the aggregated estimator (TMLE-*Ia*).

Under independent errors and minimal covariate interference (equation (22)), TMLE-*II*, constructed under the restricted causal model, performs well with good confidence interval coverage and results in notably more power (34%). However, with stronger covariate interference (equation (23)), TMLE-*II* is biased and provides misleading inference. Its confidence interval coverage is much less than nominal (81%), while deceivingly providing the most power (65%). Under the null, we also see inflated Type I error rates of 18% (Supplementary Table S2).

When the unmeasured factors determining the outcome U_Y are correlated, the assumptions in the restricted causal model (equation (11)) do not hold (Supplementary Figure S2b). As expected, the unadjusted estimator is again biased with 95% confidence interval coverage ranging from 48% to 67%. Both targeted estimators developed under the general model (TMLE-*Ia* and TMLE-*Ib*) have negligible bias, but the cluster-level estimator yields more power. The individual-level TMLE developed under the restricted model (TMLE-*II*) now exhibits substantial bias regardless of the strength of covariate interference. Its resulting confidence interval coverage is much less than the nominal and type I error reaches > 40% (Supplementary Table S2).

In summary, when the assumptions in the more restrictive causal model hold, the individual-level targeted estimator (TMLE-*II*) is the most powerful. However, if these commonly made assumptions fail, this TMLE is biased and can yield misleading inference in an observational setting. Incorporating working assumptions during the estimation stage (equation (10)) is more robust than assuming they hold in the underlying causal model (equation (11)). Specifically, TMLE-*Ib* provides a mechanism to leverage the pairing of individual-level covariates and outcomes, while avoiding additional causal assumptions. In practice, we recommend considering a general TMLE-*I* which includes both cluster-level and individual-level specifications in the Super Learner library for initial estimation of the outcome regression. This TMLE is implemented in the following simulation study.

6.2   Simulation 2 – HIV prevention and treatment trial

We now consider a more complicated simulation, generated to reflect the running example. For 1000 iterations, we simulate a cluster randomized Test-and-Treat trial, consisting of 32 communities with 200 individuals each. Within each community, we generate an underlying sexual network through a degree-corrected, bipartite stochastic block model.[69] On each network, we simulate an HIV epidemic with a susceptible-infected-recovered compartmental model.[70] In the intervention arm, 85% of the HIV-positive patients are on ART and have successfully suppressed viral replication. In the control arm, 55% of the HIV-positive patients are on ART and are suppressed.[45,71–73] There is no sexual mixing or spillover effects across communities. To initiate the epidemic in each community, we randomly select 10% of individuals to be infected and allow the virus to spread until an average prevalence

of 25% is reached. We then begin the study and follow all communities for three years. Full Python code to generate the networks and epidemic is available in Staples.[74]

As before, the target of inference is the population average treatment effect: the expected difference in the counterfactual cumulative HIV incidence under the Test-and-Treat intervention and under the standard of care. Within each community, 75 baseline HIV-negative individuals are selected, and the cluster-level outcome is the proportion who seroconvert within the three years of follow-up. The true value of the treatment effect is calculated by averaging the difference in the cluster-level counterfactual outcomes in the population of all clusters from all trials (32×1000). The estimated impact of the Test-and-Treat intervention is -4.0%, reducing HIV incidence from 9.1% under the standard of care to 5.1% under the intervention. We also simulate under the null by setting the counterfactual outcome under the intervention equal to the counterfactual outcome under the control.

We consider the following individual-level adjustment variables: demographic risk group, degree (number of sexual partners), and number of partners infected at baseline. We also consider the following cluster-level adjustment variables: baseline HIV prevalence, assortativity (degree-degree correlation across all network connections), and number of components (number of distinct sexual groups). To select among candidate adjustment variables, we apply a discrete Super Learner to data-adaptively select the candidate TMLE, which minimizes variance and maximizes precision.[59] This procedure incorporates "collaborative"[75] estimation of the known propensity score $g_0^c(A \mid E, \mathbf{W}) = g_0(A \mid E, W) = 0.5$ for further gains in precision.

We implement this approach under the larger general model (TMLE-*I*) and under the smaller sub-model (TMLE-*II*). Both TMLEs include pooled individual-level regressions as candidate estimators of the conditional mean of the cluster-level outcome $\bar{Q}_0^c(A, E, \mathbf{W})$. The former estimates the propensity score and targets at the cluster-level, while the latter estimates the propensity score and targets at the individual-level. In other words, TMLE-I can be considered a hybrid of TMLE-*Ia* (aggregated data approach) and TMLE-*Ib* (incorporating working assumptions), which were studied in the previous section. In this simulation, the restricted causal model (equation (11)) does not hold due to causal interactions between individuals within a community (i.e. sexual transmission of HIV through the network). Nonetheless, the finite sample performances of the TMLEs are expected to be similar due to randomization of the exposure (i.e. the double robustness property). We compare the targeted approaches to the unadjusted estimator, inverse probability of treatment weighting (IPTW) adjusting for average degree in the propensity score regression, and G-computation adjusting for average degree in the outcome regression.

6.2.1 Results—As expected, all estimators are unbiased and adjustment for baseline covariates increases precision and power in this trial setting (Table 3).[49,56,59,76–79] The unadjusted difference in cluster-level mean outcomes yields 66% power, while IPTW yields 68%, and parametric G-computation yields 75%. The two TMLEs, data-adaptively adjusting for the covariate(s) to increase precision, obtain substantially more power (82–83%), while maintaining nominal confidence interval coverage and Type I error control. In both TMLEs,

the number of partners infected at baseline (an individual-level covariate) is selected as the adjustment variable for the outcome regression in 76% of the trials (Supplementary Table S3). The slight difference in performance between the two TMLEs is due to targeting, which occurs at the cluster-level in TMLE-I and at the individual-level in TMLE-II. Overall, these simulations demonstrate that in a trial setting, the utility of the working assumption (equation (10)) is wider flexibility in covariate adjustment to increase efficiency without creating bias.

7 Application – household socioeconomic status and baseline HIV testing in SEARCH

The Sustainable East Africa Research in Community Health Study (SEARCH) is an ongoing cluster randomized trial to evaluate the impact of a community-based strategy for early HIV diagnosis with immediate and streamlined ART on HIV incidence in rural Uganda and Kenya (NCT:01864603). In SEARCH, population-based HIV testing was conducted through multi-disease community health campaigns, consisting of out-offacility health fairs followed by home-based testing for non-attendees.[80] HIV testing was successfully completed for 89% (131,307/146,906) of residents who were aged 15 years and considered stable (6 months in the community during the past year) at baseline. Since data collection for the primary outcome is ongoing, we apply the proposed methods to estimate the association of household socioeconomic status on the risk of not testing for HIV at baseline.

In this application, the cluster is the household, and the cluster-based exposure is an indicator of living in a household in the lowest socioeconomic class, calculated using principal component analysis of ownership of livestock and household items.[80] The individual-level outcome is an indicator of failing to test for HIV, and the cluster-level outcome is the proportion of adults not testing in a given household. The cluster-level confounders include community indicators, the size of the household, and an indicator of male head of household (Table 4). The individual-level confounders include age, sex, educational attainment, occupation type, marital status, and mobility (indicator of living 1 or more months away from the community). The target parameter is the standardized risk difference, corresponding to the causal risk difference if the necessary assumptions hold.

In this setting, we are willing to assume that after controlling for the cluster-level confounders and exposure, each individual's outcome is not a direct function of other household members' individual-level covariates. We are also willing to assume that the conditional expectation of the individual-level outcome is common across individuals. Therefore, under the general model $\mathscr{M}^I$ and using the working assumptions in equation (10), we implement TMLE with Super Learner to fully leverage the pairing of individual-level risk factors and outcomes, while avoiding unwarranted assumptions (SuperLearner-v2.0-21[81]). The library of candidate algorithms includes both parametric and semi-parametric approaches: main terms logistic regression without and without all possible pairwise interactions, generalized additive models (gam-v1.14[82]), and penalized maximum likelihood (glmnet-v2.0-5[83]). We use the same library for estimation of the outcome regression and the propensity score. The analysis is restricted to the 16 intervention

communities (77,525 adults total), and the household is the unit of independence: $J =$ 32,024.

After controlling for measured confounders, the marginal risk of not testing associated with living in household in the lowest socioeconomic class is 10.7%, while the marginal risk of not testing associated with living in a household in a higher socioeconomic class is 10.0%. Despite the large sample size, the standardized risk difference of 0.7% (95%CI: −0.1%, 1.4%) is not significant at the 0.05-level. For comparison, the unadjusted estimator, which fails to control for confounding, yielded a risk difference of −0.3% (−1.0%, 0.3%).

8   Concluding remarks

In this manuscript, we present two distinct approaches for leveraging a hierarchical data structure to improve the performance of double robust TMLEs for the causal effect of a cluster-level exposure. The first assumes a general hierarchical causal model, which allows for arbitrary dependence of individuals within clusters. For the corresponding statistical model $\mathcal{M}^I$, we review a cluster-level TMLE, which is a direct analog for the individual-level TMLE in non-hierarchical setting. Our novel contribution to this cluster-level estimator is to use the pairing of individual-level covariates and outcomes for improved estimation of the expected cluster-level outcome. Pooled individual-level regressions can lead to both asymptotic and finite sample improvements without placing restrictions on the original statistical model. Super Learner provides one way to choose between and combine several candidate algorithms, including cluster-level parametric regressions, averages of individual-level regressions, and more data-adaptive methods.

We then consider a more restrictive causal sub-model, which assumes that the cluster-level and individual i-specific covariates are sufficient to control for confounding. For the corresponding restricted statistical model $\mathcal{M}^{II}$, we present an alternative individual-level TMLE, which still targets the relevant cluster-level causal effect. When the assumptions in the sub-model hold, this TMLE is guaranteed asymptotically to be at least as efficient as the TMLE developed under the general causal model. When the assumptions fail, this TMLE may be subject to bias and misleading inference in an observational setting. However, if the propensity score is consistently estimated, the individual-level TMLE will remain consistent due to its double robustness property, representing an important advantage over alternative estimators, such as those based on a single regression (e.g. IPTW and G-computation).

The results of this paper have the following practical implications. When the exposure is delivered at the cluster-level, care should be taken when specifying the causal model and framing the statistical estimation problem. In particular, researchers need to consider if an individual's outcome could be impacted by another's covariates and if the cluster-level and individual i-specific covariates are sufficient to control for confounding. If so, the individual-level TMLE, developed under the sub-model (Figure 2), can offer asymptotic and finite sample improvements. If not, estimation under the sub-model can result in misleading inference in an observational setting. Instead, the cluster-level TMLE, developed under general model (Figure 1), is appropriate and can still harness the pairing of individual-level risk factors and outcomes. Overall, incorporating working assumptions during estimation is

more robust than assuming they hold in the underlying causal model. For both TMLEs, the use of data-adaptive estimators, such as Super Learner, avoids the parametric modeling assumptions inherent in common multilevel approaches (e.g. random effects and GEE) and improves our chances for reliable inference.

There are several areas of future work. Examples include extensions for missingness on the outcome vector, longitudinal settings, and more complicated schemes for sampling individuals within a cluster (e.g. case-control sampling). We plan to contrast the algorithms proposed in this manuscript with the two-stage TMLE, where an individual-level TMLE is used to obtain the optimal estimate of the cluster-level outcome $\hat{Y}^c$ (potentially accounting for informative measurement and missingness at the individual-level), and then a cluster-level TMLE (using these cluster-level outcomes $\hat{Y}^c$) implemented to estimate the effect of the cluster-based exposure.[84] We also plan to contrast the proposed algorithms with augmented-GEE[13,42] when the cluster size is informative.[23] Finally, we plan to generalize the proposed algorithms to estimate the effects of individual-level exposures in an infectious disease setting (e.g. vaccine studies).[20,21,85] In all cases, the hierarchical causal models presented in this manuscript ensure that the parameter of interest is defined separately from the estimation approach and reflects the underlying scientific question. This is a distinct advantage of the Targeted Learning framework over other approaches that rely on parametric regressions to define the quantity estimated and thus the scientific question answered.[17]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

1. Kawachi I and Berkman L. Neighborhoods and health. Oxford, NY: Oxford University Press, 2003.

2. Oakes J The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology (with discussion). Soc Sci Med 2004; 58: 1929–1952. [PubMed: 15020009]

3. Sobel M What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. J Am Stat Assoc Asoc 2006; 101: 1398–1407.

4. Hayes R and Moulton L. Cluster randomised trials. Boca Raton, FL: Chapman & Hall/CRC, 2009.

5. University of California, San Francisco. Sustainable East Africa Research in Community Health (SEARCH). ClinicalTrials.gov, http://clinicaltrials.gov/show/NCT01864603 (2013).

6. University of Pittsburg. Trial of a middle school coach gender violence prevention program. ClinicalTrials.gov, https://clinicaltrials.gov/show/NCT02331238 (2015).

7. Washington University School of Medicine. Staph household intervention for eradication (SHINE). ClinicalTrials.gov, https://clinicaltrials.gov/show/NCT02572791 (2015).

8. Galbraith S, Daniel J and Vissel B. A study of clustered data and approaches to its analysis. J Neurosci 2010; 30: 10601–10608. [PubMed: 20702692]

9. Laird N and Ware J. Random-effects models for longitudinal data. Biometrics 1982; 38: 963–974. [PubMed: 7168798]

10. Liang K and Zeger S. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13–22 http://www.jstor.org/stable/2336267.

11. Gardiner J, Luo Z and Roman L. Fixed effects, random effects and GEE: what are the differences? Stat Med 2009; 28: 221–239. [PubMed: 19012297]

12. Hubbard A, Ahern J, Fleischer N, et al. To GEE or not to GEE comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology 2010; 21: 467–474. [PubMed: 20220526]

13. Stephens A, Tchetgen Tchetgen E and DeGruttola V. Augmented generalized estimating equations for improving efficiency and validity of estimation in cluster randomized trials by leveraging cluster-level and individual-level covariates. Stat Med 2012; 31: 915–930. [PubMed: 22359361]

14. Stephens A, Tchetgen Tchetgen E and DeGruttola V. Locally efficient estimation of marginal treatment effects when outcomes are correlated: is the prize worth the chase? Int J Biostat 2014; 10: 59–75. [PubMed: 24566369]

15. Pearl J Probabilistic reasoning in intelligent systems. San Mateo, CA: Morgan Kaufmann, 1988.

16. Pearl J Causality: models, reasoning and inference, 2nd ed. New York, NY: Cambridge University Press, 2009.

17. van der Laan M and Rose S Targetedlearning: causal inference for observational and experimental data. London:Springer, 2011.

18. Cox D Planning of experiments. New York, NY: Wiley, 1958.

19. Rubin D Randomization analysis of experimental data: the fisher randomization test comment. J Am Stat Assoc 1980; 75: 591–593.

20. Halloran M and Struchiner C. Study designs for dependent happenings. Epidemiology 1991; 2: 331–338. [PubMed: 1742381]

21. Halloran M and Struchiner C. Causal inference in infectious diseases. Epidemiology 1995; 6: 142–151. [PubMed: 7742400]

22. Hudgens M and Halloran M. Toward causal inference with interference. J Am Stat Assoc 2008; 103: 832–842. [PubMed: 19081744]

23. Seaman S, Pavlou M and Copas A. Review of methods for handling confounding by cluster and informative cluster size in clustered data. Stat Med 2014; 33: 5371–5387. [PubMed: 25087978]

24. Rosenbaum P and Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41–55.

25. Abadie A and Imbens G. Large sample properties of matching estimators for average treatment effects. Econometrica 2006; 74: 235–267.

26. Multivariate Sekhon J. and propensity score matching software with automated balance optimization: the matching package for R. J Stat Software 2011; 42: 1–52.

27. Robins J A new approach to causal inference in mortality studies with sustained exposure periods–application to control of the healthy worker survivor effect. Math Model 1986; 7: 1393–1512.

28. Taubman S, Robins J, Mittleman M, et al. Intervening on risk factors for coronary heart disease: an application of the parametric G-formula. Int J Epidemiol 2009; 38: 1599–1611. [PubMed: 19389875]

29. Young J, Cain L, Robins J, et al. Comparative effectiveness of dynamic treatment regimes: An application of the parametric g-formula. Stat Biosci 2011; 3: 119–143. [PubMed: 24039638]

30. Snowden J, Rose S and Mortimer K. Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. Am J Epidemiol 2011; 173: 731–738. [PubMed: 21415029]

31. Horvitz D and Thompson D. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 1952; 47: 663–685.

32. Robins J, Hernán M and Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000; 11: 550–560. [PubMed: 10955408]

33. Hernán M, Brumback B and Robins J. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000; 11: 561–570. [PubMed: 10955409]

34. Bodnar L, Davidian M, Siega-Riz A, et al. Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. Am J Epidemiol 2004; 159: 926–934. [PubMed: 15128604]

35. Herna´n M and Robins J. Estimating causal effects from epidemiological data. J Epidemiol Commun Health 2006; 60: 578–586.

36. Shen C, Li X and Li L. Inverse probability weighting for covariate adjustment in randomized studies. Stat Med 2014; 33: 555–568. [PubMed: 24038458]

37. Robins J and Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers In: Jewell NP, Dietz K and Farewell VT (eds) AIDS epidemiology. Basel: Birkhäuser, 1992.

38. Robins J, Rotnitzky A and Zhao L. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 1994; 89: 846–866.

39. van der Laan M and Robins J. Unified methods for censored longitudinal data and causality. Heidelberg: Springer-Verlag, 2003.

40. Bang H and Robins J. Doubly robust estimation in missing data and causal inference models. Biometrics 2005; 61: 962–972. [PubMed: 16401269]

41. van der Laan M and Rubin D. Targeted maximum likelihood learning. Int J Biostat 2006; 2: Article 11.

42. Prague M, Wang R, Stephens A, et al. Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes. Biometrics 2016; 72: 1066–1077. [PubMed: 27060877]

43. Cohen M, Chen YQ, McCauley M, et al. Prevention of HIV-1 infection with early antiretroviral therapy. New Engl J Med 2011; 365: 493–505. [PubMed: 21767103]

44. Joint United Nations Programme on HIV/AIDS (UNAIDS). 90-90-90 an ambitious treatment target to help end the AIDS epidemic, http://www.unaids.org/en/resources/documents/2014/90-90-90 (2014).

45. World Health Organization. Guideline on when to start antiretrovial therapy and on pre-exposure prophylaxis for HIV, http://www.who.int/hiv/pub/guidelines/earlyrelease-arv/en/ (2015).

46. The TEMPRANO ANRS 12136 Study Group. A trial of early antiretrovirals and isoniazid preventive therapy in Africa. New Engl J Med 2015; 373: 808–822. [PubMed: 26193126]

47. The INSIGHT START Study Group. Initiation of antiretroviral therapy in early asymptomatic HIV infection. New Engl J Med 2015; 373: 795–807. [PubMed: 26192873]

48. Cohen M, Chen Y, Mccauley M, et al. Final results of the HPTN 052 randomized controlled trial: antiretroviral therapy prevents HIV transmission. J Acquired Immune Deficiency Syndrome 2015; 18: 20479.

49. Moore K and van der Laan M. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. Stat Med 2009; 28: 39–64. [PubMed: 18985634]

50. Gruber S and van der Laan M tmle: An R package for targeted maximum likelihood estimation. J Stat Software 2012; 51: 1–35.

51. Lendle S, Schwab J, Petersen M, et al. ltmle: An R package implementing targeted minimum loss-based estimation for longitudinal data. J Stat Software 2017; 81: 1–21.

52. Gruber S and van der Laan M. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. Int J Biostat 2010; 6: 1–14. DOI: 10.2202/1557-4679.1260.

53. Rose S and van der Laan M. Understanding TMLE In: van der Laan M and Rose S (eds) Targeted learning: causal inference for observational and experimental data. New York, NY: Springer, 2011.

54. van der Laan M and Starmans R. Entering the era of data science: targeted learning and the integration of statistics and computational data analysis. Adv Stat 2014; 14: 502678.

55. Gruber S and van der Laan M. Targeted minimum loss based estimator that outperforms a given estimator. Int J Biostat 2012; 8: 1–20.

56. Rosenblum M and van der Laan M. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. Int J Biostat 2010; 6: 1–41. DOI:10.2202/1557-4679.1138.

57. Bickel P, Klaassen C, Ritov Y, et al. Efficient and adaptive estimation for semiparametric models. New York, NY: Springer, 1997.

58. van der Laan M and Robins J. Unified methods for censored longitudinal data and causality. New York, NY: Springer-Verlag, 2003.

59. Balzer L, van der Laan M, Petersen M, et al. Adaptive pre-specification in randomized trials with and without pairmatching. Stat Med 2016; 35: 4528–4545. [PubMed: 27436797]

60. van der Laan M, Polley E and Hubbard A. Super learner. Stat Appl Genetics Mol Biol 2007; 6: 25.

61. Polley E, Rose S and van der Laan M. Super learner In: van der Laan M and Rose S (eds) Targeted learning: causal inference for observational and experimental data. New York, NY: Springer, 2011.

62. Wolpert DH. Stacked generalization. Neural Network 1992; 5: 241–259.

63. Breiman L Heuristics of instability and stabilization in model selection. Ann Stat 1996; 24: 2350–2383.

64. Cortes C and Vapnik V. Support-vector networks. Mach Learn 1995; 20: 273–297.

65. Hastie T and Tibshirani R. Generalized additive models. London: Chapman & Hall, 1990.

66. Tibshirani R Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 1996; 58: 267–288.

67. Schnitzer M, van der Laan M, Moodie E, et al. Effect of breastfeeding on gastrointestinal infection in infants: a targeted maximum likelihood approach for clustered longitudinal data. Ann Appl Stat 2014; 8: 703–725. [PubMed: 25505499]

68. R Core Team R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistica Computing 2017, http://www.R-project.org.

69. Karrer B and Newman M. Stochastic block models and community structure in networks. Phys Rev E 2011; 83: 016107.

70. Anderson R and May R. Infectious diseases of humans: dynamics and control. New York, NY: Oxford University Press, 1992.

71. Gaolathe T, Wirth K, Holme M, et al. Botswana's progress toward achieving the 2020 UNAIDS 90-90-90 antiretroviral therapy and virological suppression goals: a population-based survey. Lancet HIV 2016; 3: e221–e230. DOI: 10.1016/S2352-3018(16)00037-0. [PubMed: 27126489]

72. Fidler S, Floyd S, Yang B, et al. Towards the second UNAIDS target: Population-level ART coverage in HPTN 071 (PopART). In: Conference on retroviruses and opportunistic infections (CROI), Boston, MA, 22–25 2 2016.

73. Petersen M, Balzer L, Kwarsiima D, et al. SEARCH test and treat study in Uganda and Kenya exceeds the UNAIDs 90-90-90 cascade target by achieving over 80% population-level viral suppression after 2 years. In: 21st International AIDS conference, Durban, South Africa, 18–22 7 2016.

74. Staples P Python code for targeted PrEP in test-and-treat trials. Technical report, Harvard TH Chan School of Public Health, https://github.com/ctphoenix/HIV-PrEP-Simulation (2016).

75. Gruber S and van der Laan M. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. Int J Biostat 2010; 6: 1–29.

76. Fisher R Statistical methods for research workers, 4th ed. Edinburgh: Oliver and Boyd Ltd., 1932.

77. Cochran W Analysis of covariance: its nature and uses. Biometrics 1957; 13: 261–281.

78. Cox D and McCullagh P. Some aspects of analysis of covariance. Biometrics 1982; 38: 541–561. [PubMed: 7171689]

79. Tsiatis A, Davidian M, Zhang M, et al. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Stat Med 2008; 27: 4658–4677. [PubMed: 17960577]

80. Chamie G, Clark T, Kabami J, et al. A hybrid mobile HIV testing approach for population-wide HIV testing in rural East Africa. Lancet HIV 2016; 3: e111–e119. [PubMed: 26939734]

81. Polley E, LeDell E, Kennedy C et al. SuperLearner: Super Learner prediction, http://CRAN.R-project.org/package=SuperLearner. R package version 2.0-22 (2017).

82. Hastie T gam: Generalized additive models, http://CRAN.R-project.org/package=gam. R package version 1.14 (2016).

83. Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Software 2010; 33: 1–22, http://CRAN.R-project.org/package=glmnet. R package version 1.1-5.

84. Balzer L, Petersen M, Schwab J, et al. Estimating the impact of community-level interventions: the SEARCH trial and HIV prevention in Sub-Saharan Africa. In: Western North American Region (WNAR) of the International Biometric Society (IBS) conference, 6 2012 Fort Collins, CO.

85. Morozova O, Cohen T and Crawford F. Risk ratios for contagious outcomes. J R Soc Interface 2018; 15(20170696): 1–12.
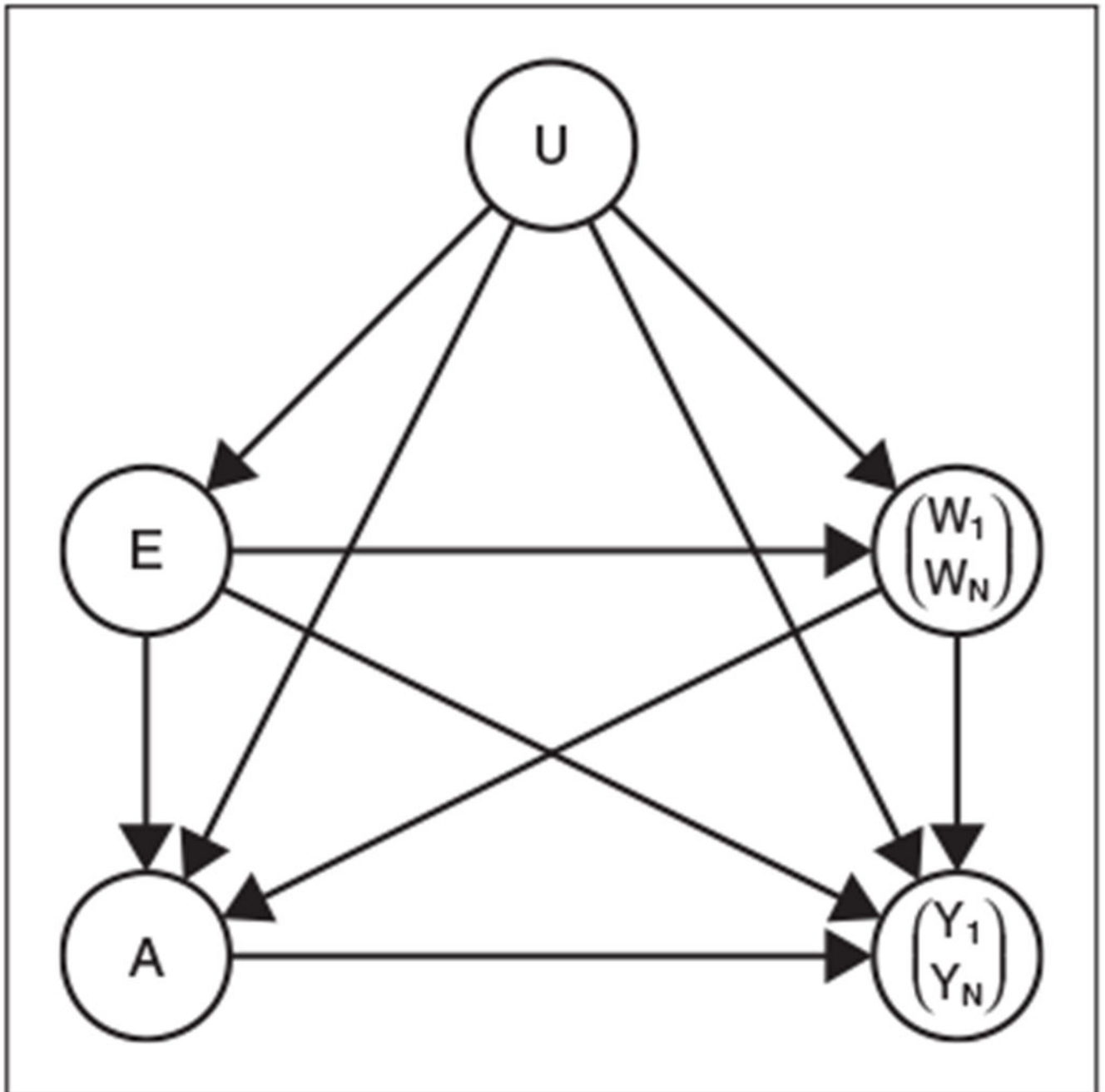
Figure 1.
Directed acyclic graph for a general hierarchical causal model (equation (1)) with U as unmeasured factors, E as cluster-level covariates, $(W_1, \ldots, W_N)$ as individual-level covariates, A as the cluster-level exposure, and $(Y_1, \ldots, Y_N)$ as individual-level outcomes. Identifiability (Section 2.3) will require additional assumptions on the unmeasured factors U (Supplementary Figure S1).
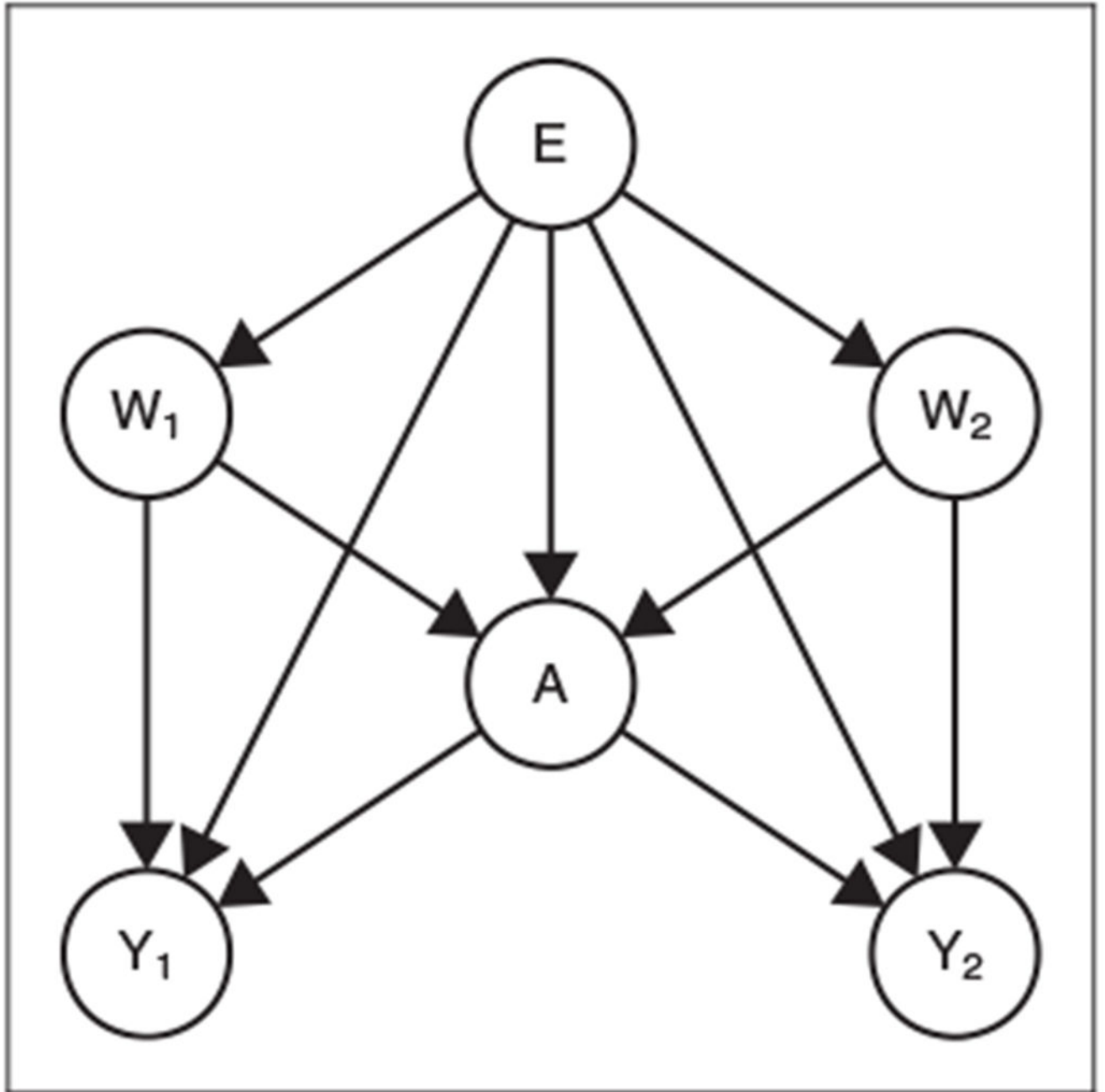
Figure 2.
Simplified directed acyclic graph for the restricted hierarchical causal model (equation (11)).
For ease of presentation, we only show two individuals, denoted by subscripts 1 and 2, in a
given cluster and assume all unmeasured factors U are independent. For additional details,
see Supplementary Figure S2.

Table 1.

Targeted estimators considered for Simulation 1: "Causal Model" refers to the causal model assumed during development of the estimator.

Estimator	Causal Model	Outcome Regression	Pscore Regression	Targeting
TMLE-*Ia*	General (equation (I))	Cluster-level	Cluster-level	Cluster-level
TMLE-*Ib*[a]	General (equation (I))	Individual-level[b]	Cluster-level	Cluster-level
TMLE-*II*	Restrictive (equation (II))	Individual-level	Individual-level	Individual-level

Note: "Cluster-level" refers to logistic regression after all the data are aggregated. "Individual-level" refers to logistic regression pooling individuals across clusters and with weights $\alpha_{ij} = 1/N_j$.

[a]During estimation consider working assumptions to generate alternative estimators of $\bar{Q}_0^c$.

[b]Run a pooled individual-level regression and then average individual-level predictions within clusters.

Table 2.

Estimator performance in Simulation 1 under minimal covariate interference (equation (22)) and under stronger covariate interference (equation (23)).

Estimator	Minimal covariate interference					Stronger covariate interference				
	Bias	σ	rMSE	Power	Coverage	Bias	σ	rMSE	Power	Coverage
Unadj.	10.4	5.0	11.5	66	46	7.6	3.8	8.5	72	49
TMLE-*Ia*	0.0	1.2	1.2	28	95	0.0	1.4	1.4	34	94
TMLE-*Ib*	0.0	1.2	1.2	27	95	0.0	1.4	1.4	23	98
TMLE-*II*	0.2	1.2	1.2	34	95	1.7	1.6	2.3	65	81
Independent U_Y determining the outcome										
Unadj.	6.3	3.2	7.1	88	48	−3.6	2.4	4.3	21	67
TMLE-*Ia*	−0.0	1.3	1.3	86	94	0.0	1.7	1.7	96	94
TMLE-*Ib*	−0.0	1.3	1.3	28	100	0.0	1.8	1.8	91	98
TMLE-*II*	−4.1	2.4	4.7	5	58	−2.1	2.0	3.0	56	81
Dependent U_Y determining the outcome										

Note: We also vary the dependence of the unmeasured factors determining the outcome $U_\mathbf{y}$: independent (top) and correlated (bottom).

Performance is given by bias as the average deviation between the estimate and truth; σ as the standard error; rMSE as the root-mean squared error; power as the proportion of times the false null hypothesis is rejected, and coverage as the proportion of times the 95% confidence interval contains the true value. All measures are percentages.

Table 3.

Estimator performance in Simulation 2 when there is an effect and under the null.

	With an effect					Under the null				
	Bias	**σ**	**rMSE**	**Power**	**Coverage**	**Bias**	**σ**	**rMSE**	**Type I**	**Coverage**
Unadj.	0.1	1.6	1.6	66	94	0.1	1.9	1.9	4	96
IPTW	0.1	1.6	1.6	68	96	0.0	1.8	1.8	3	97
Gcomp.	0.1	1.5	1.6	75	93	0.0	1.8	1.8	6	94
TMLE-*I*	0.1	1.3	1.3	83	96	0.0	1.5	1.5	5	95
TMLE-*II*	0.1	1.3	1.3	82	95	0.0	1.5	1.5	5	95

Note: Performance is measured by bias as the average deviation between the estimate and truth; σ as the standard error; rMSE as the root-mean squared error; power as the proportion of times the false null hypothesis is rejected; coverage as the proportion of times the 95% confidence interval contains the true value, and Type I error as the proportion of times the true null hypothesis is rejected. All measures are in percentage.

Table 4.

Characteristics of baseline adult residents of the 16 SEARCH intervention communities (five in Eastern Uganda, five in Southwestern Uganda, and six in Kenya) with complete socioeconomic information (249 individuals excluded).

	E. Uganda	S.W Uganda	Kenya	Overall
N. individuals	25041	24913	27571	77525
N. households	10106	9939	11979	32024
Male	11365 (45%)	11641 (47%)	12137 (44%)	35143 (45%)
Age in years				
15–24	9572 (38%)	8466 (34%)	9226 (33%)	27264 (35%)
25–34	5305 (21%)	5709 (23%)	6669 (24%)	17683 (23%)
35–44	3986 (16%)	4363 (18%)	4235 (15%)	12584 (16%)
45+	6178 (25%)	6375 (26%)	7441 (27%)	19994 (26%)
Education				
Less than primary	3855 (15%)	4413 (18%)	2132 (8%)	10400 (13%)
Primary	15255 (61%)	13966 (56%)	22302 (81%)	51523 (66%)
Secondary or higher	5931 (24%)	6534 (26%)	3137 (11%)	15602 (20%)
Occupation				
Formal[a]	5826 (23%)	5273 (21%)	6604 (24%)	17703 (23%)
High risk informal[b]	397 (2%)	652 (3%)	2331 (8%)	3380 (4%)
Low risk informal[c]	17190 (69%)	16318 (65%)	15361 (56%)	48869 (63%)
Jobless or disabled	751 (3%)	1132 (5%)	2066 (7%)	3949 (5%)
Other	877 (4%)	1538 (6%)	1209 (4%)	3624 (5%)
Never married	6913 (28%)	7424 (30%)	7515 (27%)	21852 (28%)
Mobile[d]	3024 (12%)	3305 (13%)	1960 (7%)	8289 (11%)
Male household head	18219 (73%)	16247 (65%)	16120 (58%)	50586 (65%)
Household size[e]	3 (2, 4)	3 (2, 4)	3 (2, 4)	3 (2, 4)
Lowest SES[f]	4201 (17%)	5212 (21%)	2522 (9%)	11935 (15%)
Did not test for HIV	2434 (10%)	2604 (10%)	3439 (12%)	8477 (11%)

Note: Analyses also adjusted for community indicators.

[a]Formal: teacher, student, government worker, military worker, health worker, factory worker.

[b]High risk informal: fishmonger, fisherman, bar owner, bar worker, transport, tourism.

[c]Low risk informal: farmer, shopkeeper, market vendor, hotel worker, housewife, household. worker, construction worker, mining.

[d]Mobile: 1 month/past year away from the community.

[e]Median with interquartile range.

[f]Lowest SES: Living in a household with the lowest quintile of the wealth index.