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

Laura B. Balzer*, Wenjing Zheng, Mark J. van der Laan,
Maya L. Petersen, and the SEARCH Collaboration

* Corresponding author; lbalzer@umass.edu

March 12, 2018

Appendix A - Concrete example of the general causal model

Consider the HIV prevention and treatment study. The general causal model (Eq. 2.1 and Figure 1 in the main text) describes the following data generating experiment. First the unmeasured factors U are drawn from $\mathbb{P}_U$. Informally, we can think of generating these background factors U when we sample the cluster from the target population and select individuals from that cluster. Then the community-level covariates E (e.g. region, baseline HIV prevalence, perceived need) are generated by some deterministic, but unspecified, function f_E of background factors U_E. Next the matrix of individual-level covariates $\mathbf{W}$ (e.g. demographic characteristics and risk behavior) is generated as some function $f_{\mathbf{W}}$ of the cluster-level covariates E and matrix of individual-level background factors $U_{\mathbf{W}}$. This causal model specifies that the intervention A may have been allocated among communities differentially and may depend on the cluster-level characteristics E, the matrix of individual-level characteristics $\mathbf{W}$, as well as the unmeasured factors included in U_A. Finally, this model assumes that these pre-intervention community and individual-level characteristics $(E, \mathbf{W})$ together with the intervention and unmeasured factors $(A, U_{\mathbf{Y}})$ can affect whether each individual becomes infected with HIV by the end of the study $\mathbf{Y}$.

Appendix B - Pooled individual-level causal effect

When the number of sampled individuals is constant ($N_j = n \ \forall j$), we can rewrite the treatment-specific mean as

$$\mathbb{E}\big[Y^c(a)\big] = \mathbb{E}\left[\sum_{i=1}^{n} \alpha_{i\cdot} Y_{i\cdot}(a)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y_{i\cdot}(a)\right]$$

where we have used our choice of weights $\alpha_{ij} = 1/n$. In this case, the causal effect of the cluster-based exposure on the cluster-level outcome equals the average causal effect of the cluster-based exposure on the i^{th} individual's outcome:

$$\mathbb{E}\big[Y^c(1) - Y^c(0)\big] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[Y_{i\cdot}(1) - Y_{i\cdot}(0)\big]. \tag{A.1}$$

Further, when the index i is non-informative (i.e. corresponds with the i^{th} element of a random permutation of $\{1, \ldots, n\}$), then the marginal distributions of the baseline covariates and counterfactual outcomes $(W_{i\cdot}, Y_{i\cdot}(1), Y_{i\cdot}(0))$ are constant in i. In this case, the right-hand side of equation (A.1) does not depend on i and simplifies to $\mathbb{E}\big[Y(1) - Y(0)\big]$: the expected difference in the individual-level counterfactual outcomes if all clusters received the treatment versus control level of the intervention. The expectation is now over the target population of pooled individuals from all clusters. Applied to the HIV example, this causal parameter (Eq. A.1) evaluates the difference in the risk (probability) of HIV acquisition for a randomly selected individual if all communities implemented the Test-and-Treat strategy versus if all communities continued with the standard of care.

If the number of individuals varies across clusters ($N_j \neq n \ \forall j$), then the pooled individual-level causal effect can still be defined through an alternative cluster-level outcome with weights as $\alpha_{ij} = J/\sum_j N_j$. When cluster size is informative (i.e. when the intervention effect depends on the cluster size [1]), the pooled individual-level causal effect (Eq. A.1) will generally not equal the cluster-level causal effect ($\mathbb{E}\big[Y^c(1)\big] - \mathbb{E}\big[Y^c(0)\big]$). Depending on the application, either or both may be of primary interest.

Appendix C - Additional details on loss functions

As an initial estimator of the conditional mean outcome, we can simply regress the cluster-level outcome Y^c onto the exposure and covariates $(A, E, \mathbf{W})$. We could, for example, use the squared error loss function

$$\mathcal{L}^c_{MSE}(\bar{Q}^c)(O) = \big[Y^c - \bar{Q}^c(A, E, \mathbf{W})\big]^2.$$

Alternatively, if the cluster-level outcome Y^c is standardized so that $Y^c \in (0, 1)$, then we could also use the binary log-likelihood loss function[2]:

$$-\mathcal{L}^c_{ll}(\bar{Q}^c)(O) = Y^c \log\big[\bar{Q}^c(A, E, \mathbf{W})\big] + (1 - Y^c)\log\big[1 - \bar{Q}^c(A, E, \mathbf{W})\big].$$

These regressions would result in a cluster-level analysis. For example in a linear regression model, the fitted regression parameters are defined as the least squares estimator:

$$\hat{\beta} = \arg\min_{\beta} \sum_{j=1}^{J} \big[Y^c_j - \bar{Q}^c_{\beta}(A_j, E_j, \mathbf{W}_j)\big]^2.$$

Without making additional assumptions, these loss functions can also be specified at the individual-level. For the squared error loss, we have

$$\mathcal{L}_{MSE}(\bar{Q}^c)(O) = \sum_{i=1}^{N} \alpha_{i\cdot}\big[Y_{i\cdot} - \bar{Q}^c(A, E, \mathbf{W})\big]^2$$

This is a valid loss function: $\bar{Q}^c_0 = \arg\min_{\bar{Q}^c} \mathbb{P}_0 \mathcal{L}_{MSE}(\bar{Q}^c)$. A similar result can be proved for the binary log-likelihood loss function. These loss functions would result in an individual-level regression analysis. For example in a linear regression model, the fitted regression parameters are defined as the least squares estimator:

$$\hat{\beta} = \arg\min_{\beta} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \alpha_{ij}\big[Y_{ij} - \bar{Q}^c_{\beta}(A_j, E_j, \mathbf{W}_j)\big]^2,$$

where, for example, $\alpha_{ij} = 1/N_j$. The least squares estimator $\hat{\beta}$ solves the estimating equation:

$$\begin{aligned} 0 &= \sum_{j=1}^{J}\sum_{i=1}^{N_j} \alpha_{ij} \frac{d}{d\beta}\bar{Q}^c_\beta(A_j, E_j, \mathbf{W}_j)\big(Y_{ij} - \bar{Q}^c_\beta(A_j, E_j, \mathbf{W}_j)\big) \\ &= \sum_{j=1}^{J} \frac{d}{d\beta}\bar{Q}^c_\beta(A_j, E_j, \mathbf{W}_j) \left(\sum_{i=1}^{N_j} \alpha_{ij}(Y_{ij} - \bar{Q}^c_\beta(A_j, E_j, \mathbf{W}_j)) \right). \end{aligned}$$

From this latter equation, it follows that the least squares estimator for the individual-level analysis is identical to the cluster-level least squares estimator.

Under the working model assumptions (Eq. 3.7), the squared-error loss function for $\bar{Q}_0(A, E, W) \equiv \mathbb{E}_0(Y|A, E, W)$ is now given by

$$\mathcal{L}_{MSE}(\bar{Q})(O) = \sum_{i=1}^{N} \alpha_{i.}(Y_{i.} - \bar{Q}(A, E, W_{i.}))^2.$$

A similar representation can be written for the log-likelihood loss. These loss functions would result in an individual-level regression analysis, but now with paired individual-level data $(Y_{i.}, W_{i.})$ and a much smaller adjustment set. For example in a linear regression model, the fitted regression parameters are defined as the least squares estimator:

$$\hat{\beta} = \arg\min_{\beta} \sum_{j=1}^{J}\sum_{i=1}^{N_j} \alpha_{ij}(Y_{ij} - \bar{Q}_\beta(A_j, E_j, W_{ij}))^2.$$

where, for example, $\alpha_{ij} = 1/N_j$. Thus, we could now apply Super Learner based on this loss function to estimate the common conditional mean function $\bar{Q}_0$, which then yields a fit of the object of interest $\bar{Q}^c_0(A, E, \mathbf{W}) = \sum_i \alpha_{i.}\bar{Q}_0(A, E, W_{i.})$. Assuming such a working model (Eq. 3.7) represents reality, an estimator of $\bar{Q}^c_0$ based on a pooled individual-level regression analysis may be more accurate than a cluster-level analysis, which is unable to pair individual-level outcomes and covariates.

Appendix D - Step-by-step implementation and `R` code

With hierarchical data, the cluster-level TMLE for $\Psi^I(\mathbb{P}_0)$ can be implemented in the following steps:

1. Estimate the expected cluster-level outcome given the exposure and covariates $\bar{Q}^c_0(A, E, \mathbf{W})$ using Super Learner where the library includes both cluster-level regressions and averages of individual-level regressions and where selection is based on a cluster-level loss function.

2. Use the resulting estimator $\hat{\bar{Q}}^c$ to calculate the predicted outcomes $\hat{\bar{Q}}^c(A_j, E_j, \mathbf{W}_j)$ for each cluster $j = 1, \ldots, J$.

3. Estimate the cluster-level propensity score $g^c_0(a|E, \mathbf{W})$ using parametric regression or Super Learner with a cluster-level loss function.

4. Use the resulting estimator $\hat{g}^c$ to calculate a cluster-level clever covariate $\hat{H}^c_j = \frac{\mathbb{I}(A_j=a)}{\hat{g}^c(A_j|E_j,\mathbf{W}_j)}$ for each cluster $j = 1, \ldots, J$.

5. Estimate the fluctuation coefficient ϵ by running parametric logistic regression of the cluster-level outcome Y^c on the cluster-level covariate $\hat{H}^c$ with offset as $logit(\hat{\bar{Q}}^c)$.

6. Obtain targeted predictions of the cluster-level outcome as

$$\hat{\bar{Q}}^{c*}(a, E_j, \mathbf{W}_j) = logit^{-1}\left[logit[\hat{\bar{Q}}^{c}(a, E_j, \mathbf{W}_j)] + \hat{\epsilon}\hat{H}^c_j\right]$$

   for each cluster $j = 1, \ldots, J$.

7. Obtain a point estimate by taking the empirical mean of these targeted predictions across the sample of J clusters:

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

8. Construct 95% confidence intervals for the resulting TMLE as $\hat{\Psi}^I \pm 1.96 \times \frac{\hat{\sigma}}{\sqrt{J}}$ where $\hat{\sigma}^2$ is the sample variance of the estimated influence curve $\hat{D}^I(\hat{Q}^*, \hat{g}^c)$ (Eq. 3.6 in main text).

The individual-level TMLE for $\Psi^{II}(\mathbb{P}_0)(a)$ can be implemented in the following steps:

1. Estimate the expected individual-level outcome given the exposure and covariates $\bar{Q}_0(A, E, W)$ using Super Learner where the library includes parametric and data-adaptive pooled individual-level regressions and where selection is based on a individual-level loss function. If cluster size varies, include weights $\alpha_{ij} = 1/N_j$.

2. Use the resulting estimator $\hat{\bar{Q}}$ to calculate the predicted outcomes $\hat{\bar{Q}}(A_j, E_j, W_{ij})$ for each individual $i = 1, \ldots, N_j$ in cluster $j = 1, \ldots, J$.

3. Estimate the individual-level propensity score $g_0(a|E, W_{i\cdot})$ using a pooled individual-level regression of A on $(E, W_{i\cdot})$ or using more data-adaptive methods, such as Super Learner, with a individual-level loss function. If cluster size varies, include weights $\alpha_{ij} = 1/N_j$.

4. Use the resulting estimator $\hat{g}$ to calculate an individual-level clever covariate $\hat{H}_{ij} = \frac{\mathbb{I}(A_j=a)}{\hat{g}(A_j|E_j, W_{ij})}$ for each individual $i = 1, \ldots, N_j$ in cluster $j = 1, \ldots, J$.

5. Estimate the fluctuation coefficient ϵ by running pooled parametric logistic regression of the individual-level outcome $Y_{i\cdot}$ on the individual-level covariate $\hat{H}_{i\cdot}$ with offset as $logit(\hat{\bar{Q}})$. If cluster size varies, include weights $\alpha_{ij} = 1/N_j$.

6. Use the targeted estimator to obtain predictions of the individual-level outcome $Y_{i\cdot}$ given $A = a$ and covariates as

$$\hat{\bar{Q}}^*(a, E_j, W_{ij}) = logit^{-1}\left[logit[\hat{\bar{Q}}(a, E_j, W_{ij})] + \hat{\epsilon}\hat{H}_{ij}\right]$$

   for each individual i in each cluster j.

7. Obtain a point estimate by taking the empirical mean of these targeted predictions within clusters and then across the sample of J clusters:

$$\hat{\Psi}^{II}(\hat{Q}^*)(a) = \frac{1}{J}\sum_{j=1}^{J}\sum_{i=1}^{N_j} \alpha_{ij}\hat{\bar{Q}}^*(a, E_j, W_{ij}).$$

8. Construct 95% confidence intervals for the resulting TMLE as $\hat{\Psi}^{II} \pm 1.96 \times \frac{\hat{\sigma}}{\sqrt{J}}$ where $\hat{\sigma}^2$ is the sample variance of the estimated influence curve $D^{II}(\hat{Q}^*, \hat{g})$.

Full `R` code for the simulations and estimators is at `https://github.com/LauraBalzer/HierarchicalTMLE`.

Appendix E- Theoretical comparison of the TMLEs

Proof. Suppose that the true observed data distribution $\mathbb{P}_0$ is an element of the sub-model $\mathcal{M}^{II}$. Then we have $\Psi^I(\mathbb{P}_0)(a) = \Psi^{II}(\mathbb{P}_0)(a) = \psi_0(a)$. For simplicity, also consider a randomized trial with $g_0^c(A|E, \mathbf{W}) = g_0(A|E, W) = 0.5$. Then we can re-write the efficient influence curves as

$$D^I(\mathbb{P}_0)(O) = 2\mathbb{I}(A = a)\big(Y^c - \bar{Q}_0^c(A, E, \mathbf{W})\big) + \bar{Q}_0^c(a, E, \mathbf{W}) - \psi_0(a) \tag{A.2}$$

and

$$D^{II}(\mathbb{P}_0)(O) = \sum_{i=1}^{N} \big[\alpha_i.2\mathbb{I}(A = a)\big(Y_i. - \bar{Q}_0(A, E, W_i.)\big) + \bar{Q}_0(a, E, W_i.) - \psi_0(a)\big] \tag{A.3}$$

Due to the linearity of summations, one can show that in this setting $D^I(\mathbb{P}_0)(O) = D^{II}(\mathbb{P}_0)(O)$ and thus the efficiency bound is the same. □

Supplementary figures

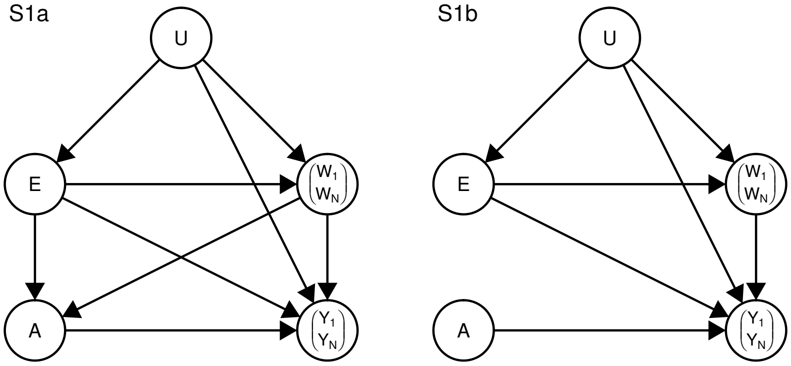
S1a
U
E
$\begin{pmatrix}W_1\\W_N\end{pmatrix}$
A
$\begin{pmatrix}Y_1\\Y_N\end{pmatrix}$
S1b
U
E
$\begin{pmatrix}W_1\\W_N\end{pmatrix}$
A
$\begin{pmatrix}Y_1\\Y_N\end{pmatrix}$

Figure S1: Two possible directed acyclic graphs (DAGs) that are compatible with the no unmeasured confounders assumption in the general causal model. Here, U denotes unmeasured factors, E the cluster-level covariates, $(W_{1.}, \dots W_{N.})$ the individual-level covariates, A the cluster-level exposure, and $(Y_{1.}, \dots, Y_{N.})$ the individual-level outcomes. ***S1a:*** *an observational setting where the covariates $(E, (W_{1.}, \dots W_{N.}))$ are sufficient to control for confounding.* ***S1b:*** *cluster randomized trial where by design there is no confounding.*

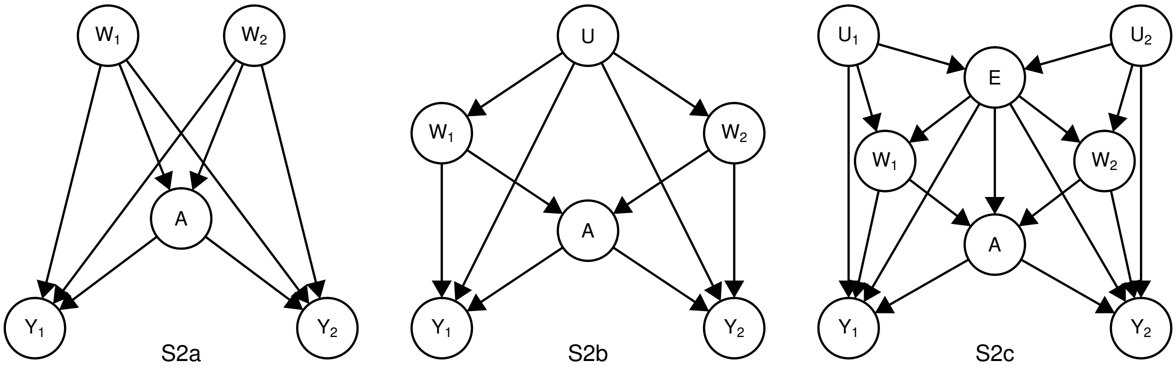

Figure S2: Directed acyclic graphs (DAGs) to illustrate the assumptions on the distribution of unmeasured factors. Let U denote unmeasured factors, E the cluster-level covariates, W the individual-level covariates, A the cluster-level exposure, and Y the individual-level outcome. For ease of presentation, we only show two individuals, denoted by subscripts 1 and 2, in a given cluster. In all causal models, the measured covariates capture all the common causes of the exposure and outcomes. ***S2a:*** *For simplicity, we ignore the cluster-level covariates E. Even if all the unmeasured factors are independent (and thus not explicitly shown), we need to control for both $(W_{1.}, W_{2.})$ when there is covariate interference (i.e $Y_{1.}$ is a function of $W_{2.}$ and $Y_{2.}$ is a function of $W_{1.}$). The assumptions in the restricted causal model do not hold.* ***S2b:*** *For simplicity, we again ignore the cluster-level covariates E. Even with no covariate interference, we need to control for both $(W_{1.}, W_{2.})$ when there is a shared unmeasured common cause of the individual-level covariates and individual-level outcomes. The assumptions in the restricted causal model do not hold.* ***S2c:*** *Let $U_{1.}$ and $U_{2.}$ denote the i-specific unmeasured common causes of the cluster-level covariates, individual-level covariates, and individual-level outcome. Even with no covariate interference, we need to control for $(E, W_{1.}, W_{2.})$, because the cluster-level covariates E are a collider of the $U_{1.}$ and $U_{2.}$. The assumptions in the restricted causal model do not hold.*

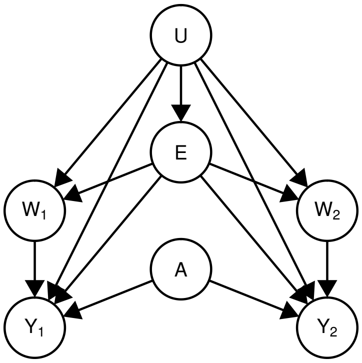

Figure S3: When the cluster-level exposure is randomized, we do not need to adjust for covariates, regardless of the error structure. If there is also no covariate interference, the assumptions in the restricted causal model do hold.

Supplementary tables

References

[1] S.R. Seaman, M. Pavlou, and A.J. Copas. Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in Medicine*, 33:5371–5387, 2014.

[2] S. Gruber and M.J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1):Article 26, 2010. doi: 10.2202/1557-4679.1260.

*Supplementary Table S1: True value of the causal effect of the cluster-level exposure $\mathbb{E}[Y^c(1) - Y^c(0)]$ for each of the data generating processes in Simulation 1. When there is a treatment effect, the coefficient for the exposure in the logistic regression for the conditional probability of the **individual-level** outcome (Eq. 6.1-6.2) is 0.1. Nonetheless, the strength of the effect of the cluster-level exposure on the **cluster-level** outcome depends on the presence or absence of strong covariate interference as well as the presence or absence of dependence in the unmeasured factors determining the individual-level outcomes $U_{\mathbf{Y}}$. By construction, the treatment effect is always 0 in the null setting. All measures are in %.*

	With an effect		Under the Null	
	Indpt. $U_{\mathbf{Y}}$	Dept. $U_{\mathbf{Y}}$	Indpt. $U_{\mathbf{Y}}$	Dept. $U_{\mathbf{Y}}$
Minimal covariate interference	1.6	3.8	0	0
Stronger covariate interference	2.1	6.3	0	0

Supplementary Table S2: Estimator performance in Simulation 1 under minimal covariate interference (Eq. 6.1) and under stronger covariate interference (Eq. 6.2). We also vary the dependence of the unmeasured factors determining the individual-level outcomes: 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; type I error as the proportion of times the true null hypothesis is rejected, and coverage as the proportion of times the 95% confidence interval contains the true value. All measures are in %.

	Minimal covariate interference					Stronger covariate interference				
Estimator	Bias	σ	rMSE	Type I	Coverage	Bias	$\hat{\sigma}$	rMSE	Type I	Coverage
Unadj.	10.4	5.1	11.5	54	46	7.6	3.9	8.5	51	49
TMLE-Ia	-0.0	1.2	1.2	6	94	-0.0	1.4	1.4	6	94
TMLE-Ib	-0.0	1.2	1.2	5	95	-0.0	1.4	1.4	2	98
TMLE-II	0.2	1.2	1.2	6	94	1.6	1.6	2.3	18	82
Independent $U_{\mathbf{Y}}$ determining the outcome										
Unadj.	6.5	3.3	7.3	53	47	-3.8	2.5	4.5	34	66
TMLE-Ia	-0.0	1.3	1.3	5	95	0.0	1.8	1.8	6	94
TMLE-Ib	-0.0	1.3	1.3	0	100	0.0	1.8	1.8	2	98
TMLE-II	-4.2	2.3	4.8	43	57	-2.3	2.1	3.1	19	81
Dependent $U_{\mathbf{Y}}$ determining the outcome										

Supplementary Table S3: For the TMLEs developed under the general model $\mathcal{M}^I$ and under the sub-model $\mathcal{M}^{II}$, the number of times a candidate variable was selected for adjustment during initial estimation of the outcome regression or the known propensity score in Simulation 2. The candidates include nothing ("Unadj."), degree, demographic risk group ("Demo."), the number of partners infected at baseline ("N. partners"), cluster-level baseline HIV prevalence, assortativity ("Assort."), and the number of distinct sexual groups ("N. components").

	Unadj.	Degree	Demo.	N. partners	Prevalence	Assort.	N. components
Selection under the general model (TMLE-$\mathcal{M}^I$)							
Outcome regression	2	64	4	759	112	8	51
Propensity	830	36	38	8	33	25	30
Selection under the sub-model (TMLE-$\mathcal{M}^{II}$)							
Outcome regression	2	64	4	759	112	8	51
Propensity score	877	14	6	8	33	26	36