Estimating the impact of community-level interventions: The SEARCH Trial and HIV Prevention in Sub-Saharan Africa

Laura Balzer, University of California, Berkeley
Maya Petersen, University of California, Berkeley
Joshua Schwab, University of California, Berkeley
Mark van der Laan, University of California, Berkeley

Available at: https://works.bepress.com/laura_balzer/3/
Estimating the impact of community-level interventions: The SEARCH Trial and HIV Prevention in Sub-Saharan Africa

Laura Balzer
Maya Petersen, Joshua Schwab
Mark van der Laan

Division of Biostatistics
University of California, Berkeley
lbbalzer@berkeley.edu

Members of the SEARCH Consortium

June 2012
1. The SEARCH Trial
   - Efficient Community Cohort Design: optimize study validity
Outline

1. The SEARCH Trial
   - Efficient Community Cohort Design: optimize study validity

2. From design to estimation
   - Observed data, statistical model and target parameter
Outline

1. The SEARCH Trial
   - Efficient Community Cohort Design: optimize study validity
2. From design to estimation
   - Observed data, statistical model and target parameter
3. Estimation of the average treatment effect
   - Targeted Minimum Loss-based Estimation (TMLE): obtain unbiased and efficient estimands
   - Two-stage estimation of the community-specific outcome and then the average treatment effect
1 The SEARCH Trial
   - Efficient Community Cohort Design: optimize study validity

2 From design to estimation
   - Observed data, statistical model and target parameter

3 Estimation of the average treatment effect
   - Targeted Minimum Loss-based Estimation (TMLE): obtain unbiased and efficient estimands
   - Two-stage estimation of the community-specific outcome and then the average treatment effect

4 Simulations & Conclusion
Large community randomized trial to be conducted in Uganda and Kenya

Multinational, interdisciplinary collaboration between UCB, UCSF and our in-country partners

Study Goal: Estimate the effect of early HIV diagnosis combined with streamlined antiretroviral therapy (regardless of CD4 level) on HIV incidence in rural East African communities.
SEARCH is a five-year trial, involving 32 communities with \( \approx 5,000 \) adults each.

Annually a community health campaign (CHC) will offer HIV testing and multi-disease prevention and treatment services.

- In intervention communities, all individuals testing positive for HIV will be offered ART (regardless of CD4 levels) and linked to care.
- In control communities, all individuals testing positive for HIV will be offered ART according to in-county guidelines.

Study Hypothesis: immediate treatment and streamlined care will reduce five-year cumulative incidence of HIV.
The Efficient Community Cohort Design (ECCO): Optimizing study validity

- A baseline household census to enumerate stable residents.
- Annual community health campaigns (CHC), where attendees are linked over years with unique identifiers.
- Supplementary tracking of a random sample of individuals not attending each campaign.
  - Tested for HIV and linked to care.
With the Efficient Community Cohort Design, we

- have an enumerated cohort for effect estimation.
  - cohort corresponds to the entire community.
- incidence is measured in the same manner in control and treatment communities (i.e. CHC + tracking).
- incidence is measured directly at multiple time points.
- link individuals over successive waves of the intervention.
  - obtain longitudinal individual-level data.
- recover from loss to follow-up and informative missingness in a financially and logistically feasible manner.
Given

1. Question of interest: what is the causal effect of immediate treatment and streamlined care on the five-year cumulative incidence of HIV?

2. Study Design: Efficient Community Cohort
From Design to Estimation

Given

1. Question of interest: what is the causal effect of immediate treatment and streamlined care on the five-year cumulative incidence of HIV?

2. Study Design: Efficient Community Cohort

We are now ready to define

1. the observed data

2. the statistical model

3. the target causal parameter and target statistical parameter
For the $n$ communities, we measure
- baseline community-level characteristics: $E$
- the randomly assigned treatment: $A$
For the $n$ communities, we measure
- baseline community-level characteristics: $E$
- the randomly assigned treatment: $A$

For the $J'$ individuals within a community, we measure
- baseline covariates (measured at the census): $W(0)$
- an indicator of CHC attendance at time $t$: $\Delta^*(t)$
- an indicator of having data collected at time $t$
  (attended CHC or tracked with known probability): $\Delta(t)$
- time-varying covariates and HIV status at time $t$
  (if he/she was observed): $\Delta(t)(W(t), Y(t))$

with $t = 1, \ldots, 6$
Statistical model is

- the set of possible observed data distributions.
- only restricted by assumptions known to hold by design:
  - The treatment is randomized.
  - The missingness mechanism is known:
    \[
    \mathbb{P}(\Delta(t) = 1) = \Delta^*(t) + (1 - \Delta^*(t))p(t)
    \]
    where \( p(t) \) is the tracking probability at time \( t \)
- semi-parametric.
For simplicity, let us focus on the baseline HIV negative cohort in each community.

- Set of $J$ individuals with $Y_j(1) = 0$. 
For simplicity, let us focus on the baseline HIV negative cohort in each community.

- Set of $J$ individuals with $Y_j(1) = 0$.

Then within an enumerated HIV negative cohort, let $Y_c$ represent the five-year cumulative incidence of HIV for a given community.

- $Y_c$ is the probability of becoming HIV positive within five years.
- $Y_c$ is not directly observed as the final HIV status is unknown for some cohort members.
Ideally, we would observe the final HIV status of all cohort members under both the intervention $A = 1$ and the control $A = 0$.

Let $Y_a^c$ denote the counterfactual cumulative incidence of HIV under no censoring at the final time point and under treatment $A = a$. 
Ideally, we would observe the final HIV status of all cohort members under both the intervention $A = 1$ and the control $A = 0$.

Let $Y_a^c$ denote the counterfactual cumulative incidence of HIV under no censoring at the final time point and under treatment $A = a$.

Then the target causal parameter is

$$\Psi^F = \mathbb{E}[Y_1^c] - \mathbb{E}[Y_0^c]$$

and equals the average treatment effect under this joint intervention.
Let us first focus on a single community and the intervention to prevent censoring at the final time point.

Then given the tracking process, the cumulative incidence of HIV under this intervention can be identified...
Let us first focus on a single community and the intervention to prevent censoring at the final time point.

Then given the tracking process, the cumulative incidence of HIV under this intervention can be identified

- through inverse weighting as

\[
Y^c = \mathbb{E} \left[ \frac{\mathbb{I}(\Delta(6) = 1)}{\mathbb{P}(\Delta(6) = 1|\Delta^*(6))} Y(6) \right]
\]

- or by the G-computation formula with

\[
Y^c = \mathbb{E}\left[ \mathbb{E}[Y(6)|\Delta(6) = 1, \Delta^*(6)] \right]
\]

Adjusting for additional covariates is not necessary for identifiability, but can improve efficiency.
Given each community’s outcome $Y^c$ and the randomization of the treatment $A$, the average treatment effect can be identified through inverse weighting as:

$$\Psi_0 = E\left[I(A = 1)P(A = 1)\right] Y^c - E\left[I(A = 0)P(A = 0)\right] Y^c$$

or by the G-computation formula with:

$$\Psi_0 = E\left[Y^c | A = 1\right] - E\left[Y^c | A = 0\right]$$

Adjusting for measured covariates is not necessary for identifiability, but can improve efficiency.
From Design to Estimation: Target Statistical Parameter

Given each community’s outcome $Y^c$ and the randomization of the treatment $A$, the average treatment effect can be identified through inverse weighting as

\[ \Psi_0 = \mathbb{E} \left[ \frac{\mathbb{I}(A = 1)}{\mathbb{P}(A = 1)} Y^c \right] - \mathbb{E} \left[ \frac{\mathbb{I}(A = 0)}{\mathbb{P}(A = 0)} Y^c \right] \]

or by the G-computation formula with

\[ \Psi_0 = \mathbb{E}[Y^c|A = 1] - \mathbb{E}[Y^c|A = 0] \]

Adjusting for measured covariates is not necessary for identifiability, but can improve efficiency.
The causal parameter has been expressed in terms of the observed data.
- We considered two point treatment interventions.
- The estimation problem is defined.
Stage 1: Estimation of community-specific outcome $Y^c$

Naïve estimator:

$$
\hat{Y}^c = \frac{1}{J} \sum_j \left( \frac{I(\Delta_j(6))}{P(\Delta_j(6) = 1)} Y_j(6) \right)
$$

- Observed proportion of seroconversions
- Biased: censoring informed by HIV status
- Inefficient: ignores measured covariate data
Stage 1: Estimation of community-specific outcome $Y^c$

Inverse Probability Weighted (IPW) estimator:

$$\hat{Y}^c = \frac{1}{j} \sum_j \frac{\mathbb{I}(\Delta_j(6) = 1)}{\mathbb{P}(\Delta_j(6) = 1 | \Delta^*_j(6))} Y_j(6)$$

- correct for informative missingness by re-weighting with the conditional probability of being observed.
- Unbiased: if tracking is 100% successful.
- Inefficient: if an individual’s covariates (beyond final CHC attendance $\Delta^*(6)$) are predictive of final HIV status
  - e.g. a prior positive HIV test
Stage 1:
Estimation of community-specific outcome $Y^c$

Inverse Probability Weighted (IPW) estimator II:

$$\hat{Y}^c = \frac{1}{J} \sum_{j} \frac{\mathbb{I}(\Delta_j(6) = 1)}{P(\Delta_j(6) = 1 | P_{a_j})} Y_j(6)$$

where $P_{a}$ denotes the individual’s measured past prior to the final censoring node.

- Unbiased: if tracking is 100% successful.
- Unbiased: if tracking is less than 100% successful and the measured covariates, $P_{a}$, are sufficient to control for informative missingness.
- More efficient but not maximally
Stage 1: Estimation of community-specific outcome $Y^c$

Targeted Minimum Loss-based Estimation (TMLE):

1. Requires an initial estimate of the conditional mean function:

$$\bar{Q}^0(1, Pa) = \mathbb{E}[Y(6)|\Delta(6) = 1, Pa]$$

and of the treatment mechanism:

$$\mathbb{P}(\Delta(6) = 1|Pa)$$

- With a large semi-parametric model, calls for data-adaptive machine learning.
Stage 1:
Estimation of community-specific outcome $Y^c$

Targeted Minimum Loss-based Estimation (TMLE):

1. Updates the initial estimate to obtain the optimal bias-variance tradeoff for the target parameter.

- Requires a loss function

$$-L(\bar{Q}) = \log(\bar{Q})^{Y(6)} + \log(1 - \bar{Q})^{1-Y(6)}$$

and parametric sub-model through the initial estimate

$$\logit[\bar{Q}(\epsilon)] = \logit[\bar{Q}^0] + \epsilon H^*[\Delta(6), Pa]$$

such that the score at $\epsilon = 0$ spans the appropriate component of the efficient influence curve

$$D^* = \left(\frac{\mathbb{I}(\Delta(6) = 1)}{\mathbb{P}(\Delta(6) = 1|Pa)}\right) \left( Y(6) - \bar{Q}(1, Pa) \right) + \bar{Q}(1, Pa) - \mathbb{E}[\bar{Q}(1, Pa)]$$
Stage 1: Estimation of community-specific outcome $Y^c$

Targeted Minimum Loss-based Estimation (TMLE):

1. Substitutes the updated estimate, $\bar{Q}^*(\Delta(6), Pa)$, into the target parameter mapping:

$$\hat{Y}^c = \frac{1}{j} \sum_{j} \bar{Q}^*(\Delta_j(6) = 1, Pa_j)$$

- Recall the target statistical parameter:

$$Y^c = \mathbb{E}[\mathbb{E}[Y(6)|\Delta(6) = 1, Pa]]$$
Stage 1:
Estimation of community-specific outcome $Y^c$

Targeted Minimum Loss-based Estimation (TMLE):

- Unbiased: if tracking is 100% successful.
- Unbiased: if tracking is less than 100% successful and the measured covariates, $P_a$, are sufficient to control for informative missingness.
- Maximally efficient: solves the efficient score equation.
- Other properties: well-defined, substitution estimator, double robust, asymptotically normal.
Two-Stage Estimation

- Discussed several estimators of the community-level outcome $Y^c$.
- Given the estimated cumulative incidence $\hat{Y}^c$ for each community, we are ready to move to stage two: estimating the average treatment effect.
Stage 2: Estimation of the average treatment effect

Naïve estimator:

$$\hat{\Psi} = \frac{1}{n} \sum_{i} \left( \frac{\mathbb{I}(A_i = 1)}{\mathbb{P}(A_i = 1)} \hat{Y}_i^c \right) - \frac{1}{n} \sum_{i} \left( \frac{\mathbb{I}(A_i = 0)}{\mathbb{P}(A_i = 0)} \hat{Y}_i^c \right)$$

- Simple difference in treatment-specific means.
- Unbiased: treatment is randomized
- Inefficient: ignores measured covariate data
Stage 2: Estimation of the average treatment effect

Targeted Minimum Loss-based Estimation (TMLE):

1. Requires an initial estimate of the conditional mean function:

\[ \tilde{Q}^0(A, E^c) = \mathbb{E}[Y^c | A, E^c] \]

and of the treatment mechanism:

\[ P(A | E^c) = P(A) \]

- with \( E^c \) denoting the baseline community-level covariates (including summary measures of pre-treatment individual-level covariates)
Stage 2: Estimation of the average treatment effect

Targeted Minimum Loss-based Estimation (TMLE):

- Updates the initial estimate to obtain the optimal bias-variance tradeoff for the target parameter.
  - Requires a loss function and parametric sub-model through the initial estimate such that the score at $\epsilon = 0$ spans the efficient influence curve.
  - In practice, run a simple logistic regression of the outcome on the treatment and covariates.
Stage 2:
Estimation of the average treatment effect

Targeted Minimum Loss-based Estimation (TMLE):

- Substitutes the updated estimate, $\bar{Q}^*(A, E^c)$, into the target parameter mapping:

$$
\hat{\Psi} = \frac{1}{n} \sum_i (\bar{Q}^*(1, E^c_i) - \bar{Q}^*(0, E^c_i))
$$

- Recall the target statistical parameter:

$$
\Psi_0 = \mathbb{E}(Y^c|A = 1) - \mathbb{E}(Y^c|A = 0)
= \mathbb{E}\left[\mathbb{E}(Y^c|A = 1, E^c) - \mathbb{E}(Y^c|A = 0, E^c)\right]
$$
Stage 2: Estimation of the average treatment effect

Targeted Minimum Loss-based Estimation (TMLE):

- Unbiased: treatment is randomized
- Maximally efficient: solves the efficient score equation.
- Other properties: well-defined, substitution estimator, double robust, asymptotically normal.
Simulations to compare bias and variance of the two-stage estimators.

- Four estimators of the cumulative incidence of HIV in each community \(Y^c\):
  - Naïve, IPW, IPW*, TMLE
Simulations to compare bias and variance of the two-stage estimators.

- Four estimators of the cumulative incidence of HIV in each community $Y^c$:
  - Naïve, IPW, IPW*, TMLE
- Given the estimated $\hat{Y}^c$, two estimators of the average treatment effect $\Psi$:
  - Naïve, TMLE
Simulations

- The SEARCH team plans to sample 32 communities each with 4000 HIV- adults at baseline.
- Suppose that \( \sim 75\% \) of the cohort members attend the CHC annually.
- Further suppose that a random sample of 10\% of the non-attendees are tracked and tested for HIV annually.
- For these simulations,
  - The five-year cumulative incidence of HIV was 0.0178 in the treated communities and 0.0295 in the control communities.
  - The average treatment effect was \( \Psi = -0.0117 \).
Simulations: Bias over 100 iterations
ECCO to optimize study validity

<table>
<thead>
<tr>
<th></th>
<th>Naïve</th>
<th>IPW</th>
<th>IPW*</th>
<th>TMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>0.00185</td>
<td>6.03e-05</td>
<td>-1.89e-05</td>
<td>7.02e-05</td>
</tr>
<tr>
<td>TMLE</td>
<td>0.00188</td>
<td>5.18e-05</td>
<td>-1.05e-05</td>
<td>6.29e-05</td>
</tr>
</tbody>
</table>

- Columns: Stage 1: Estimation of the community-level outcome $Y^c$
- Rows: Stage 2: Estimation of the average treatment effect $\Psi$

The naïve estimator is over an order of magnitude more biased than the other estimators, which adjust for informative missingness.
Simulations: Bias over 100 iterations
ECCO to optimize study validity

Suppose there was no tracking and an unmeasured confounder affected both CHC attendance and HIV status.

<table>
<thead>
<tr>
<th></th>
<th>Naïve</th>
<th>IPW</th>
<th>IPW*</th>
<th>TMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>0.00223</td>
<td>-</td>
<td>0.00579</td>
<td>0.000175</td>
</tr>
<tr>
<td>TMLE</td>
<td>0.00228</td>
<td>-</td>
<td>0.00586</td>
<td>0.000193</td>
</tr>
</tbody>
</table>

- Columns: Stage 1: Estimation of the community-level outcome $Y^c$
- Rows: Stage 2: Estimation of the average treatment effect $\Psi$

All estimators are biased. The Efficient Community Cohort design is essential to study validity.
Suppose a random sample of 10% of non-attendees were selected for tracking, but only 60% could be successfully located and tested.

<table>
<thead>
<tr>
<th></th>
<th>Naïve</th>
<th>IPW</th>
<th>IPW*</th>
<th>TMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>0.00202</td>
<td>0.00117</td>
<td>0.001360</td>
<td>0.000269</td>
</tr>
<tr>
<td>TMLE</td>
<td>0.00206</td>
<td>0.00119</td>
<td>0.001440</td>
<td>0.000292</td>
</tr>
</tbody>
</table>

- **Columns**: Stage 1: Estimation of the community-level outcome $Y^c$
- **Rows**: Stage 2: Estimation of the average treatment effect $\Psi$

Adjusting for measured covariates can help recover from informative censoring.
Simulations: St. Dev. over 100 iterations
TMLE for efficiency

Let us again suppose a random 10% of the non-attendees were tracked with 100% success.

<table>
<thead>
<tr>
<th></th>
<th>Naïve</th>
<th>IPW</th>
<th>IPW*</th>
<th>TMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>-</td>
<td>0.00329</td>
<td>0.00300</td>
<td>0.00300</td>
</tr>
<tr>
<td>TMLE</td>
<td>-</td>
<td>0.00193</td>
<td>0.00150</td>
<td>0.00144</td>
</tr>
</tbody>
</table>

- Columns: Stage 1: Estimation of the community-level outcome $Y^c$
- Rows: Stage 2: Estimation of the average treatment effect $\Psi$

By adjusting for individual-level covariates, a more efficient estimate of the community-specific outcome is obtained.
Let us again suppose a random 10% of the non-attendees were tracked with 100% success.

<table>
<thead>
<tr>
<th></th>
<th>Naïve</th>
<th>IPW</th>
<th>IPW*</th>
<th>TMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>-</td>
<td>0.00329</td>
<td>0.00300</td>
<td>0.00300</td>
</tr>
<tr>
<td>TMLE</td>
<td>-</td>
<td>0.00193</td>
<td>0.00150</td>
<td>0.00144</td>
</tr>
</tbody>
</table>

- Columns: Stage 1: Estimation of the community-level outcome $Y^c$
- Rows: Stage 2: Estimation of the average treatment effect $\Psi$

By adjusting for community-level covariates, a more efficient estimate of the average treatment effect is obtained.
Let us again suppose a random 10% of the non-attendees were tracked with 100% success.

<table>
<thead>
<tr>
<th></th>
<th>Naïve</th>
<th>IPW</th>
<th>IPW*</th>
<th>TMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>-</td>
<td>0.00329</td>
<td>0.00300</td>
<td>0.00300</td>
</tr>
<tr>
<td>TMLE</td>
<td>-</td>
<td>0.00193</td>
<td>0.00150</td>
<td>0.00144</td>
</tr>
</tbody>
</table>

- **Columns:** Stage 1: Estimation of the community-level outcome $Y^c$
- **Rows:** Stage 2: Estimation of the average treatment effect $\Psi$

The lowest variance is achieved using TMLE for the community-specific outcome and then the average treatment effect.
The SEARCH trial will quantify the effect of immediate streamlined treatment of all HIV+ individuals on community health, economic and educational productivity.

- Presented a novel study design to optimize study validity.
- Presented a novel TMLE to obtain unbiased and efficient estimates.
- Incorporating individual and community covariates helps
  - to identify of the target parameter if tracking is incomplete.
  - to achieve effect estimates with greater precision.
Acknowledgements & the SEARCH Team

UCSF:

Principle Investigator:
Diane Havlir, MD

Investigators:
Edwin Charlebois, MPH, PhD
Craig Cohen, MD, MPH
Elvin Geng, MD, MPH
Gabriel Chami, MD, MPH
Vivek Jain, MD, MAS
Tamara Clark, MHS
Gertrude Khumalo-Sakutukwa, MSc, MMed Sc
Doug Black, BA
James G. Kahn, MD, MPH
Elliot Marseille, PhD, MPP

MU-UCSF Research Collaboration:
Co-Principle Investigator:
Moses Kamya, MBChB, MMed, MPH, PhD

Investigators:
Jane Achan, MBChB, MPed
Jane Kabami, MPH
Catherine Tugaineyo, MBA
Bridget Nzaruara, MBChB
Jennifer Namusobya, MBChB, MPH
Gideon Amanyire, MBChB, MPH
Dalsone Kwarisiima, MBChB, MPH
Dathan Byonanebye, MBChB
Geoff Lavoy, MCDBA

Kenya Medical Research Institute:

Investigators:
Elizabeth Bukusi, MBChB, MMed, MPH, PhD
Zachary Kwena, MA
Barrack Onyango, CPA, MBA

University of North Carolina, Chapel Hill and World Bank:

Investigators:
Harsha Thirumurthy, PhD