Leveraging the causal effects of stochastic interventions to evaluate vaccine efficacy in two-phase trials

Nima Hejazi

Division of Biostatistics, and Center for Computational Biology, University of California, Berkeley

@nshejazi
@nhejazi
nimahejazi.org

with D. Benkeser, M. van der Laan, H. Janes, P. Gilbert

SER: “Methods for the thorny challenges of real studies”
The burden of HIV-1

- The HIV-1 epidemic — the facts:
  - now in its fourth decade,
  - 2.5 million new infections occurring annually worldwide,
  - new infections outpace patients starting antiretroviral therapy.

- *Most efficacious* preventive vaccine: 31% reduction rate.

- **Question:** To what extent can HIV-1 vaccines be improved by modulating immunogenic CD4+/CD8+ response profiles?
HVTN 505 trial examined new antibody boost vaccines

- HIV Vaccine Trials Network’s (HVTN) 505 vaccine efficacy; randomized controlled trial, $n = 2504$ (Hammer et al. 2013).

- Immunogenic response profiles only available for two-phase sample of $n = 189$ (Janes et al. 2017) due to cost limitations.

- Two-phased sampling mechanism: 100% inclusion rate if HIV-1 positive in week 28; based on matching otherwise.

- **Question**: How would HIV-1 infection risk in week 28 have changed had immunogenic response (due to vaccine) differed?
Two-phase sampling censors the complete data structure

- Complete (unobserved) data $X = (L, A, Y) \sim P^X_0 \in \mathcal{M}^X$, as per the full HVTN 505 trial cohort (Hammer et al. 2013):
  - $L$ (baseline covariates): sex, age, BMI, behavioral HIV risk;
  - $A$ (exposure): immunogenic response profiles (CD4+, CD8+);
  - $Y$ (outcome of interest): HIV-1 infection status at week 28.

- Observed data $O = (C, CX) = (L, C, CA, Y); C \in \{0, 1\}$ is an indicator for inclusion in the two-phase sample.

- Can we use the two-phase sample ($n = 189$) to estimate causal effects in the vaccine arm ($n \approx 1400$)? How?
Stochastic interventions define the causal effects of shifts

- Causal estimand: counterfactual mean of HIV-1 infection under a *shifted* immunogenic response distribution.

- Díaz and van der Laan (2012; 2018): *Shift* interventions?

\[
d(a, l) = \begin{cases} 
  a + \delta, & \text{if plausible} \\
  a, & \text{otherwise}
\end{cases}
\]

- Díaz and van der Laan (2012; 2018) give a statistical target parameter and influence function for the complete data case.
HIV-1 risk under shifted immunogenic responses

Immunogenic response profile

Mean risk of HIV-1 infection
HIV-1 risk under shifted immunogenic responses

**Immunogenic response profile**

**Mean risk of HIV–1 infection**
HIV-1 risk under shifted immunogenic responses
HIV-1 risk under shifted immunogenic responses

Immunogenic response profile

Mean risk of HIV–1 infection
HIV-1 risk under shifted immunogenic responses

**Immunogenic response profile**

**Mean risk of HIV–1 infection**
HIV-1 risk under shifted immunogenic responses

Immunogenic response profile

Mean risk of HIV–1 infection
HIV-1 risk under shifted immunogenic responses

**Immunogenic response profile**

- Vertical axis: 0 to 150
- Horizontal axis: -2 to 2

**Mean risk of HIV-1 infection**

- Vertical axis: 0.00 to 1.00
- Horizontal axis: None
HIV-1 risk under shifted immunogenic responses

Immunogenic response profile

Mean risk of HIV−1 infection
HIV-1 risk under shifted immunogenic responses

![Immunogenic response profile](chart1)

![Mean risk of HIV-1 infection](chart2)
HIV-1 risk under shifted immunogenic responses

Immunogenic response profile

Mean risk of HIV–1 infection
Efficient estimators in spite of two-phase sampling

- What if sampling mechanism \( \pi_0(Y, L) = \mathbb{P}(C = 1 \mid Y, L) \) is not known by design? Nonparametric estimation of \( \pi_0(Y, L) \)?

- Building on Rose and van der Laan (2011), we provide
  - asymptotically linear and nonparametric-efficient estimators;
  - multiply robust, with two forms of double robustness;
  - Gaussian limit distributions and Wald-type confidence intervals.

- New open source software for easily using these estimators:
  - https://github.com/nhejazi/haldensify (densities)
  - https://github.com/nhejazi/txshift (one-step, TMLE)
Fighting the HIV-1 epidemic (Hejazi et al. 2020)

Figure 1: Analysis of HIV-1 risk as a function of CD8+ immunogenicity, using R package txshift (https://github.com/nhejazi/txshift.)
We can target immunogenic responses modulated by HIV-1 vaccines to improve future efficacy against HIV-1.

*Stochastic* interventions constitute a flexible framework for considering *realistic* intervention policies.

Large-scale vaccine trials often use two-phase designs — need to (carefully!) adjust for sampling complications.

We’ve developed open source software for assessing the causal effects of stochastic interventions in two-phase designs.
Thank you!

https://nimahejazi.org

https://twitter.com/nshejazi

https://github.com/nhejazi

https://doi.org/10.1111/biom.13375
Appendix
Assumption 1: **Consistency**

\[ Y_{d(a_i,l_i)}^d = Y_i \text{ in the event } A_i = d(a_i,l_i), \text{ for } i = 1, \ldots, n \]

Assumption 2: **SUTVA**

\[ Y_{d(a_i,l_i)}^d \text{ does not depend on } d(a_j,l_j) \text{ for } i = 1, \ldots, n \text{ and } j \neq i, \text{ or lack of interference (Rubin 1978; 1980)} \]

Assumption 3: **Strong ignorability**

\[ A_i \perp \perp Y_{i}^{d(a_i,l_i)} | L_i, \text{ for } i = 1, \ldots, n \]
Assumption 4: Positivity (or overlap)

\[ a_i \in A \implies d(a_i, l_i) \in A \text{ for all } l \in L, \text{ where } A \text{ denotes the support of } A \text{ conditional on } L = l_i \text{ for all } i = 1, \ldots, n \]

- This positivity assumption is not quite the same as that required for categorical interventions.
- In particular, we do not require that the intervention density place mass across all strata defined by \( L \).
- Rather, we merely require the post-intervention quantity be seen in the observed data for given \( a_i \in A \) and \( l_i \in L \).
NPSEM with static interventions

- Use a nonparametric structural equation model (NPSEM) to describe the generation of \( X \) (Pearl 2009), specifically

\[
L = f_L(U_L); \ A = f_A(L, U_A); \ Y = f_Y(A, L, U_Y)
\]

- Implies a model for the distribution of counterfactual random variables generated by interventions on the process.

- A static intervention replaces \( f_A \) with a specific value \( a \) in its conditional support \( A \mid L \).

- This requires specifying a particular value of the exposure under which to evaluate the outcome \( a \) priori.
**NPSEM with stochastic interventions**

- *Stochastic interventions* modify the value $A$ would naturally assume by drawing from a modified exposure distribution.

- Consider the post-intervention value $A^* \sim G^*(\cdot \mid L)$; static interventions are a special case (degenerate distribution).

- Such an intervention generates a counterfactual random variable $Y_{G^*} := f_Y(A^*, L, U_Y)$, with distribution $P^\delta_0$.

- We aim to estimate $\psi_{0,\delta} := \mathbb{E}_{P^\delta_0}\{Y_{G^*}\}$, the counterfactual mean under the post-intervention exposure distribution $G^*$. 
Díaz and van der Laan (2012; 2018)’s *stochastic* interventions

\[ d(a, l) = \begin{cases} 
   a + \delta, & a + \delta < u(l) \quad \text{(if plausible)} \\
   a, & a + \delta \geq u(l) \quad \text{(otherwise)} 
\end{cases} \]

- Our estimand is \( \psi_{0,d} := \mathbb{E}_{P_d} \{ Y_{d(A,L)} \} \), mean of \( Y_{d(A,L)} \).

- Statistical target parameter is \( \Psi(P_{0X}) = \mathbb{E}_{P_{0X}} \bar{Q}(d(A, L), L) \), counterfactual mean of the *shifted* outcome mechanism.

- For HVTN 505, \( \psi_{0,d} \) is the counterfactual risk of HIV-1 infection, had the observed value of the immune response been altered under the rule \( d(A, L) \) defining \( G^*(\cdot \mid L) \).
Proposal: Evaluate outcome under an altered intervention distribution — e.g., $P_\delta(g_0)(A = a \mid L) = g_0(a - \delta(L) \mid L)$.

Identification conditions for a statistical parameter of the counterfactual outcome $\psi_{0,d}$ under such an intervention.

Show that the causal quantity of interest $\mathbb{E}_0\{Y_{d(A,L)}\}$ is identified by a functional of the distribution of $X$:

$$
\psi_{0,d} = \int_L \int_A \mathbb{E}_{P_0^X}\{Y \mid A = d(a, l), L = l\} \cdot q_{0,A}(a \mid L = l) \cdot q_{0,L}(l) d\mu(a) d\nu(l)
$$

Provides a derivation based on the efficient influence function (EIF) with respect to the nonparametric model $\mathcal{M}$. 
• *Proposal:* Characterization of stochastic interventions as *modified treatment policies* (MTPs).

• Assumption of *piecewise smooth invertibility* allows for the intervention distribution of any MTP to be recovered:

\[
g_{0,\delta}(a \mid l) = \sum_{j=1}^{J(l)} I_{\delta,j} \{ h_j(a, l), I \} g_0 \{ h_j(a, l) \mid I \} h_j'(a, l)
\]

• Such intervention policies account for the natural value of the intervention \( A \) directly yet are interpretable as the imposition of an altered intervention mechanism.

• Identification conditions for assessing the parameter of interest under such interventions appear technically complex (at first).
Establishes equivalence between g-formula when proposed intervention depends on natural value and when it does not.

This equivalence leads to a sufficient positivity condition for estimating the counterfactual mean under MTPs via the same statistical functional studied in Díaz and van der Laan (2012).

Extends earlier identification results, providing a way to use the same statistical functional to assess $\mathbb{E}Y_d(A, L)$ or $\mathbb{E}Y_d(L)$.

The authors also consider limits on implementing shifts $d(A, L)$, and address working in a longitudinal setting.
- Builds on the original proposal, accommodating MTP-type shifts $d(A, L)$ proposed after their earlier work.

- To protect against positivity violations, considers a specific shifting mechanism:

\[
d(a, l) = \begin{cases} 
a + \delta, & a + \delta < u(l) \\
a, & \text{otherwise}
\end{cases}
\]

- Proposes an improved “1-TMLE” algorithm, with a single auxiliary covariate for constructing the TML estimator.

- Our (first) contribution: implementation of this algorithm.
The efficient influence function (EIF) is:

\[ D(P_0^X)(x) = H(a, l)(y - \overline{Q}(a, l)) + \overline{Q}(d(a, l), l) - \psi(P_0^X). \]

The one-step estimator corrects bias by adding the empirical mean of the estimated EIF to the substitution estimator:

\[ \psi_n^+ = \frac{1}{n} \sum_{i=1}^{n} \overline{Q_n}(d(A_i, L_i), L_i) + D_n(O_i). \]

The TML estimator is built by updating initial estimates of \( \overline{Q_n} \) via a (logistic) tilting model, yielding

\[ \psi_n^* = \frac{1}{n} \sum_{i=1}^{n} \overline{Q_n^*}(d(A_i, L_i), L_i). \]

Both estimators are CAN even when nuisance parameters are estimated via flexible, machine learning techniques.
Rose and van der Laan (2011) introduce the IPCW-TMLE, to be used when observed data is subject to two-phase sampling.

*Initial proposal:* correct for two-phase sampling by using a loss function with inverse probability of censoring weights:

$$\mathcal{L}(P^X_0)(O) = \frac{C}{\pi_0(Y, L)} \mathcal{L}^F(P^X_0)(X)$$

When the sampling mechanism $\pi_0(Y, L)$ can be estimated by a parametric form, this procedure yields an efficient estimator.

However, when machine learning is used (e.g., when $\pi_0(Y, L)$ is not known by design), this is insufficient.
• Then, the IPCW augmentation must be applied to the EIF:

\[
D(P_0^X)(o) = \frac{c}{\pi_0(y, l)} D^F(P_0^X)(x) - \left(1 - \frac{c}{\pi_0(y, l)}\right).
\]

\[
\mathbb{E}(D^F(P_0^X)(x) \mid C = 1, Y = y, L = l),
\]

• Expresses observed data EIF \( D^F(P_0^X)(o) \) in terms of full data EIF \( D^F(P_0^X)(x) \); inclusion of second term ensures efficiency.

• The expectation of the full data EIF \( D^F(P_0^X)(x) \), taken only over units selected by the sampling mechanism (i.e., \( C = 1 \)).

• A unique multiple robustness property — combinations of \((g_0(L), \bar{Q}_0(A, L)) \times (\pi_0(Y, L), \mathbb{E}(D^F(P_0^X)(x) \mid C = 1, Y, L))\).
1. Construct initial estimators $g_n$ of $g_0(A, L)$ and $Q_n$ of $Q_0(A, L)$, perhaps using data-adaptive regression techniques.

2. For each observation $i$, compute an estimate $H_n(a_i, l_i)$ of the auxiliary covariate $H(a_i, l_i)$.

3. Estimate the parameter $\epsilon$ in the logistic regression model

$$\logit Q_{\epsilon,n}(a, l) = \logit Q_n(a, l) + \epsilon H_n(a, l),$$

or an alternative regression model incorporating weights.

4. Compute TML estimator $\Psi_n$ of the target parameter, defining update $Q^*_{n}$ of the initial estimate $Q_{n,\epsilon_n}$:

$$\Psi_n = \Psi(P^*_n) = \frac{1}{n} \sum_{i=1}^{n} Q^*_n(d(A_i, L_i), L_i).$$
Algorithm for IPCW-TML estimation

1. Using all observed units \((X)\), estimate sampling mechanism 
\(\pi(Y, L)\), perhaps using data-adaptive regression methods.

2. Using only observed units in the two-phase sample \(C = 1\), construct initial estimators 
\(g_n(A, L)\) and \(\bar{Q}_n(A, L)\), weighting by the sampling mechanism estimate 
\(\pi_n(Y, L)\).

3. With the approach described for the full data case, compute 
\(H_n(a_i, l_i)\), and fluctuate submodel via logistic regression.

4. Compute IPCW-TML estimator \(\Psi_n\) of the target parameter, 
by solving the IPCW-augmented EIF estimating equation.

5. Iteratively update estimated sampling weights \(\pi_n(Y, L)\) and 
IPCW-augmented EIF, updating TML estimate in each 
iteration, until 
\(\frac{1}{n} \sum_{i=1}^{n} EIF_i < \frac{1}{n}\).
Key properties of TML estimators

- **Asymptotic linearity:**

  \[ \psi(P_n^*) - \psi(P_0^X) = \frac{1}{n} \sum_{i=1}^{n} D(P_0^X)(X_i) + o_p \left( \frac{1}{\sqrt{n}} \right) \]

- **Gaussian limiting distribution:**

  \[ \sqrt{n}(\psi(P_n^*) - \psi(P_0^X)) \rightarrow N(0, \text{Var}(D(P_0^X)(X))) \]

- **Statistical inference:**

  Wald-type confidence interval: \( \psi(P_n^*) \pm z_{1 - \frac{\alpha}{2}} \cdot \frac{\sigma_n}{\sqrt{n}} \),

  where \( \sigma_n^2 \) is computed directly via \( \sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} D^2(\cdot)(X_i) \).
Identifying the best efficient estimator

Figure 2: Relative performance of reweighted and augmented estimators.
Briefly consider a simple data structure: \( X = (Y, A) \); we seek to model the outcome \( Y \) as a function of \( A \).

To posit a linear model, consider \( Y_i = \beta_0 + \beta_1 A_i + \epsilon_i \), with error \( \epsilon_i \sim N(0, 1) \).

Letting \( \delta \) be a change in \( A \), \( Y_{A+\delta} - Y_A \) may be expressed

\[
\mathbb{E} Y_{A+\delta} - \mathbb{E} Y_A = [\beta_0 + \beta_1 (\mathbb{E} A + \delta)] - [\beta_0 + \beta_1 (\mathbb{E} A)]
= \beta_0 - \beta_0 + \beta_1 \mathbb{E} A - \beta_1 \mathbb{E} A + \beta_1 \delta
= \beta_1 \delta
\]

Thus, a *unit shift* in \( A \) (i.e., \( \delta = 1 \)) may be seen as inducing a change in the difference in outcomes of magnitude \( \beta_1 \).
A causal inference perspective

- Consider a data structure: \((Y_a, a \in A)\).
- To posit a linear model, let \(Y_a = \beta_0 + \beta_1 a + \epsilon_a\) for \(a \in A\), with error \(\epsilon_a \sim N(0, \sigma_a^2)\) \(\forall a \in A\).
- For the counterfactual outcomes \((Y_{a'} + \delta, Y_{a'})\), their difference, \(Y_{a'} + \delta - Y_{a'}\), for some \(a' \in A\), may be expressed

\[
\mathbb{E}Y_{a'} + \delta - \mathbb{E}Y_{a'} = [\beta_0 + \beta_1 (a' + \delta) + \mathbb{E}\epsilon_{a'} + \delta] - [\beta_0 + \beta_1 a' + \mathbb{E}\epsilon_{a'}] = \beta_1 \delta
\]

- Thus, a *unit shift* for \(a' \in A\) (i.e., \(\delta = 1\)) may be seen as inducing a change in the difference in the counterfactual outcomes of magnitude \(\beta_1\).
Consider the stochastic intervention $g^*(\cdot \mid L)$:

$$\mathbb{E}Y_{g^*} = \int_L \int_a \mathbb{E}(Y \mid A = a, L)g(a - \delta \mid L) \cdot da \cdot dP_0(L)$$

$$= \int_L \int_z \mathbb{E}(Y \mid A = z + \delta, L)g(z \mid L) \cdot dz \cdot dP_0(L),$$

defining the change of variable $z = a - \delta$.

For a semiparametric model, $\mathbb{E}(Y \mid A = z, L) = \beta z + \theta(L)$:

$$\mathbb{E}Y_{g^*} - \mathbb{E}Y = \int_L \int_z \left[ \mathbb{E}(Y \mid A = z + \delta, L) - \mathbb{E}(Y \mid A = z, L) \right] g(z \mid L) \cdot dz \cdot dP_0(L)$$

$$= [\beta(z + \delta) + \theta(L)] - [\beta z + \theta(L)]$$

$$= \beta \delta$$
To compute the auxiliary covariate $H(a, l)$, we need to estimate conditional densities $g(A \mid L)$ and $g(A - \delta \mid L)$.

There is a rich literature on density estimation, we follow the approach proposed in Díaz and van der Laan (2011).

To build a conditional density estimator, consider

$$g_{n, \alpha}(a \mid L) = \frac{\mathbb{P}(A \in [\alpha_{t-1}, \alpha_t) \mid L)}{\alpha_t - \alpha_{t-1}},$$

for $\alpha_{t-1} \leq a < \alpha_t$.

- This is a classification problem, where we estimate the probability that a value of $A$ falls in a bin $[\alpha_{t-1}, \alpha_t)$.
- The choice of the tuning parameter $t$ corresponds roughly to the choice of bandwidth in classical kernel density estimation.
Díaz and van der Laan (2011) propose a re-formulation of this classification approach as a set of hazard regressions.

To effectively employ this proposed re-formulation, consider

\[
\mathbb{P}(A \in [\alpha_{t-1}, \alpha_t] \mid L) = \mathbb{P}(A \in [\alpha_{t-1}, \alpha_t] \mid A \geq \alpha_{t-1}, L) \times \\
\prod_{j=1}^{t-1} \left\{ 1 - \mathbb{P}(A \in [\alpha_{j-1}, \alpha_j] \mid A \geq \alpha_{j-1}, L) \right\}
\]

The likelihood of this model may be expressed to correspond to the likelihood of a binary variable in a data set expressed via a long-form repeated measures structure.

Specifically, the observation of \(X_i\) is repeated as many times as intervals \([\alpha_{t-1}, \alpha_t]\) are before the interval to which \(A_i\) belongs, and the binary variables indicating \(A_i \in [\alpha_{t-1}, \alpha_t]\) are recorded.
Density estimation with the Super Learner algorithm

- To estimate \( g(A \mid L) \) and \( g(A - \delta \mid L) \), use a pooled hazard regression, spanning the support of \( A \).

- We rely on the Super Learner algorithm of van der Laan et al. (2007) to build an ensemble learner that optimally weights each of the proposed regressions, using cross-validation (CV).

- The Super Learner algorithm uses \( V \)-fold CV to train each proposed regression model, weighting each by the inverse of its average risk across all \( V \) holdout sets.

- By using a library of regression estimators, we invoke the result of van der Laan et al. (2004), who prove this likelihood-based cross-validated estimator to be asymptotically optimal.


