Evaluating the causal impacts of vaccine-induced immune responses in two-phase vaccine efficacy trials

Nima Hejazi

Wednesday, 20 January 2021

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

@nshejazi
@nhejazi
nimahejazi.org

with M. van der Laan, H. Janes, P. Gilbert, D. Benkeser
Biostatistics seminar, Fred Hutch
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).

- Question: How would HIV-1 infection risk in week 28 have changed had vaccine-induced immunogenic response differed?

- Immunogenic response profiles only available for second-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.
Two-phase sampling censors the complete data structure

- Complete (unobserved) data $X = (L, A, S, Y) \sim P_0^X \in \mathcal{M}^X$, as per the full HVTN 505 trial cohort (Hammer et al. 2013):
  - $L$ (baseline covariates): sex, age, BMI, behavioral HIV risk,
  - $A$ (treatment): vaccination status (randomized),
  - $S$ (exposure): immune response profile for CD4+ and CD8+,
  - $Y$ (outcome of interest): HIV-1 infection status at week 28.
- Observed data $O = (C, CX) = (L, C, CS, Y)$.
  - $C \in \{0, 1\}$ indicates inclusion in the second-phase sample.
  - Implicitly conditioning on the vaccine arm, i.e., $O = X \mid A = 1$. 
NPSEM with static interventions

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

$$L = f_L(U_L); S = f_S(L, U_S); Y = f_Y(S, L, U_Y)$$

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

- A *static intervention* replaces $f_S$ with a specific value $s$ in its conditional support $S \mid A = 1, L$.

- This requires specifying a particular value of the exposure under which to evaluate the outcome *a priori*. 
- Stochastic interventions modify the value $S$ would naturally assume by drawing from a modified exposure distribution.

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

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

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

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

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

Statistical target parameter is \( \Psi(P^X_0) = \mathbb{E}_{P^X_0} Q(d(S, L), L) \), 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 \( d(S, L) \) defining \( G^*(\cdot \mid L) \).
Assumption 1: Stable Unit Treatment Value (SUTVA)

- $Y_i^{d(s_i, l_i)}$ does not depend on $d(s_j, l_j)$ for $i = 1, \ldots, n$ and $j \neq i$, or lack of interference (Rubin 1978; 1980)
- $Y_i^{d(s_i, l_i)} = Y_i$ in the event $S_i = d(s_i, l_i)$, for $i = 1, \ldots, n$

Assumption 2: Ignorability

$S_i \perp \perp Y_i^{d(s_i, l_i)} | L_i$, for $i = 1, \ldots, n$

Assumption 3: Positivity

$s_i \in S \implies d(s_i, l_i) \in S$ for all $l \in L$, where $S$ denotes the support of $S$ conditional on $L = l_i$ for all $i = 1, \ldots, n$
- **Proposal:** Evaluate outcome under an altered *intervention distribution* — e.g., \( P_\delta(g_0)(S = s \mid L) = g_0(s - \delta(L) \mid L) \).

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

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

\[
\psi_{0,d} = \int_L \int_S \mathbb{E}_{P_0^X}\{Y\mid S = d(s, l), L = l\} \cdot q_{0,S}(s \mid L = l) \cdot q_{0,L}(l) d\mu(s) d\nu(l)
\]

- Provides a derivation based on the efficient influence function (EIF) with respect to the nonparametric model \( \mathcal{M} \).
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
HIV-1 risk under shifted immunogenic responses

![Graph showing Immunogenic response profile and Mean risk of HIV-1 infection](image)

- 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

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

**Immunogenic response profile**

**Mean risk of HIV–1 infection**
The efficient influence function (EIF) is:

\[ D(P^X_0)(x) = H(s, l)(y - Q(s, l)) + Q(d(s, l), l) - \psi(P^X_0). \]

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

\[ \psi^+_n = \frac{1}{n} \sum_{i=1}^{n} Q_n(d(S_i, L_i), L_i) + D_n(O_i). \]

The TML estimator updates initial estimates of \( Q_n \) by tilting:

\[ \psi^*_n = \frac{1}{n} \sum_{i=1}^{n} Q^*_n(d(S_i, L_i), L_i). \]

Both estimators are doubly robust.
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:

\[
\mathcal{L}(P_0^X)(O) = \frac{C}{\pi_0(Y, L)} \mathcal{L}^F(P_0^X)(X)
\]

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.
Efficient estimation and multiple robustness

- Then, the IPCW augmentation must be applied to the EIF:

\[
D(P_0^X)(o) = \frac{C}{\pi_0(y, l)} D^F(P_0^X)(x) - \left(1 - \frac{C}{\pi_0(y, l)}\right) \cdot \mathbb{E}(D^F(P_0^X)(x) \mid C = 1, Y = y, L = l),
\]

- 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 \((g_0(L), \overline{Q}_0(S, L)) \times (\pi_0(Y, L), \mathbb{E}(D^F(P_0^X)(x) \mid C = 1, Y, L)).\)
Fighting the HIV-1 epidemic with preventive vaccines

Figure 1: Analysis of HIV-1 risk as a function of CD8+ immunogenicity, using R package txshift (https://github.com/nhejazi/txshift).
Big picture takeaways

- Vaccine efficacy evaluation helps to develop enhanced vaccines better informed by biological properties of the target disease.

- HIV-1 vaccines modulate immunogenic response profiles as part of their mechanism for lowering HIV-1 infection risk.

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

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

- We’ve developed robust, open source statistical software for assessing stochastic interventions in observational studies.
Thank you!

https://nimahejazi.org

https://twitter.com/nshejazi

https://github.com/nhejazi

https://doi.org/10.1111/biom.13375
At Warp Speed – COVID-19 Vaccine Trials
COVID-19 Vaccine Development

- **Nucleic acid vaccines**: Moderna (mRNA), Pfizer (mRNA)
- **Viral-vectored vaccines**: AstraZeneca (chimpanzee adenovirus), Janssen (human adenovirus)
- **Subunit vaccines**: NovaVax, Sanofi / GlaxoSmithKline
- **Weakened/inactivated vaccines**: Sinopharm, Sinovac
Operation Warp Speed (OWS)

- Do we have the time? Polio (7 years), Measles (9 years), Chickenpox (34 years), Mumps (4 years), HPV (15 years).

- OWS: “300M doses of safe, effective vaccine by 01 Jan. 2021”.

- How? Typical process timeline (73 months) replaced by an accelerated process of 14 months.

- COVID-19 Prevention Network (CoVPN):
  - formed by NIAID to establish a unified clinical trial network for evaluating vaccines and monoclonal antibodies.
  - Statisticians: primary trial design/analysis, sequential efficacy monitoring, safety monitoring, immune correlates.
Correlate of Protection (CoP): immune marker statistically predictive of vaccine efficacy, not necessarily mechanistic.

Mechanistic CoP (mCoP): immune marker that is causally and mechanistically responsible for protection.

Nonmechanistic CoP (nCoP): immune marker that is predictive but not a causal agent of protection.

A CoP is a *candidate surrogate* endpoint (Prentice 1989) — primary endpoint in future trials if reliably predictive.
Running assays on > 30,000 blood draws is timely, expensive, and, as it turns out, statistically unnecessary.

Instead we measure immune responses via a case-cohort design (Prentice 1986):
- a stratified random subcohort (≈ 1600 individuals)
- all SARS-CoV-2 and COVID endpoints.

Case-cohort designs are a special case of two-phase sampling (Breslow et al. 2003; 2009):
- Phase 1: measure baseline, vaccine, endpoint on everyone.
- Phase 2: given baseline, vaccine, endpoint, select members of immune response subcohort with (possibly known) probability.
Stochastic–Interventional Vaccine Efficacy

- Causal parameter based on vaccine efficacy (VE) estimands:
  \[
  \text{SVE}(\delta) = 1 - \frac{E[\mathbb{P}(Y = 1 \mid A = 1, S = s + \delta, L = l) \mid A = 1, L]}{\mathbb{P}(Y(0) = 1)}.
  \]

- \(\mathbb{P}(Y(0) = 1)\): counterfactual infection risk in the placebo arm — under randomization, \(\mathbb{P}(Y(0) = 1) = \mathbb{P}(Y = 1 \mid A = 0)\).

- Summarizes VE thru stochastic interventions indexed by \(\delta\).

Additional Complexities of Two-Phase Designs

- Observed data structure: \( O = (L, A, Z, CS, Y, C) \)
  - \( A \in \{0, 1\} \): randomized vaccination assignment
  - \( Z \): post-vaccination confounder (e.g., unblinded risky behavior)
  - \( S \): candidate mCoPs (causal mediators)
  - \( Y \): symptomatic SARS-CoV-2 infection
  - \( C := f(Y, L) \): selection into second-phase sample

- But what about \( O = (L, A, Z, CS, \Delta, \tilde{T}, C) \)?
  - \( \tilde{T} = \min(T_F, T_C) \): possibly right-censored time to infection
  - \( \Delta = \mathbb{I}(T_F < T_C) \): symptomatic SARS-CoV-2 infection
  - Can \( C \) still be a function of \( \tilde{T} \)?
Causal Mediation Analysis: Explanation and Mechanism

- **Identification assumptions:**
  - A1: No unmeasured confounding of \( \{A, Y\} \) relationship.
  - A2: No unmeasured confounding of \( \{S, Y\} \) relationship.
  - A3: No unmeasured confounding of \( \{A, S\} \) relationship.
  - A4: No \( \{S, Y\} \) confounder affected by \( A \), i.e., no \( Z \).

- **Indirect effects:** thru pathways involving candidate mCoPs.
  - Natural (in)direct effects (Robins and Greenland 1992, Pearl 2013): binary \( A \) and \( S \), no \( Z \), “cross-world” independence.
  - Stochastic (in)direct effects (Díaz and Hejazi 2020): continuous \( A \) and \( S \), no \( Z \); no “cross-world” exclusion.
  - Interventional (in)direct effects (Díaz et al. 2020): binary \( A \), continuous \( S \), \( Z \) OK, no “cross-world” exclusion.
  - Stochastic interventional (in)direct effects (Hejazi et al. 2020): continuous \( A \) and \( S \), \( Z \) ok, no “cross-world” exclusion.
Appendix
- **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:

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

- Such intervention policies account for the natural value of the intervention *S* 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 $E Y_d(S, L)$ or $E Y_d(L)$.

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

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

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

- Proposes an improved TMLE algorithm, with a single auxiliary covariate for constructing the TML estimator.
Nonparametric conditional density estimation

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

- 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

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

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

  - This is a classification problem, where we estimate the probability that a value of $S$ 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 reformulation of this classification approach as a set of hazard regressions.

To effectively employ this proposed reformulation, consider

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

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 $S_i$ belongs, and the binary variables indicating $S_i \in [\alpha_{t-1}, \alpha_t)$ are recorded.
To estimate \( g(S \mid L) \) and \( g(S - \delta \mid L) \), use a pooled hazard regression, spanning the support of \( S \).

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.
Key properties of TML estimators

- **Asymptotic linearity:**

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

- **Gaussian limiting distribution:**

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

- **Statistical inference:**

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

where \(\sigma_n^2\) is computed directly via \(\sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} D^2(\cdot)(X_i)\).
Algorithm for TML estimation

1. Construct initial estimators $g_n$ of $g_0(S, L)$ and $Q_n$ of $Q_0(S, L)$, perhaps using data-adaptive regression techniques.

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

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

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

   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 $Q_{n, \epsilon_n}$:

   $\Psi_n = \Psi(P^*_n) = \frac{1}{n} \sum_{i=1}^{n} Q^*_n(d(S_i, L_i), L_i).$
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 second-stage sample $C = 1$, construct initial estimators $g_n(S, L)$ and $Q_n(S, L)$, weighting by the sampling mechanism estimate $\pi_n(Y, L)$.

3. With the approach described for the full data case, compute $H_n(s_i, l_i)$, 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 $\frac{1}{n} \sum_{i=1}^{n} \text{EIF}_i < \frac{1}{n}$. 
Identifying the best efficient estimator

![Graph showing relative performance of reweighted and augmented estimators.](image)

Figure 2: Relative performance of reweighted and augmented estimators.
A linear modeling perspective

- Briefly consider a simple data structure: $X = (Y, S)$; we seek to model the outcome $Y$ as a function of $S$.
- To posit a linear model, consider $Y_i = \beta_0 + \beta_1 S_i + \epsilon_i$, with error $\epsilon_i \sim N(0, 1)$.
- Letting $\delta$ be a change in $S$, $Y_{S+\delta} - Y_S$ may be expressed

$$EY_{S+\delta} - EY_S = [\beta_0 + \beta_1 (ES + \delta)] - [\beta_0 + \beta_1 (ES)]$$

$$= \beta_0 - \beta_0 + \beta_1 ES - \beta_1 ES + \beta_1 \delta$$

$$= \beta_1 \delta$$

- Thus, a unit shift in $S$ (i.e., $\delta = 1$) may be seen as inducing a change in the difference in outcomes of magnitude $\beta_1$. 
Consider a data structure: \((Y_s, s \in S)\).

To posit a linear model, let \(Y_s = \beta_0 + \beta_1 s + \epsilon_s\) for \(s \in S\), with error \(\epsilon_s \sim N(0, \sigma_s^2) \forall s \in S\).

For the counterfactual outcomes \((Y_{s'} + \delta, Y_{s'})\), their difference, \(Y_{s'} + \delta - Y_{s'}\), for some \(s' \in S\), may be expressed:

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

Thus, a unit shift for \(s' \in S\) (i.e., \(\delta = 1\)) may be seen as inducing a change in the difference in the counterfactual outcomes of magnitude \(\beta_1\).
Slope in a semiparametric model

- Consider the stochastic intervention $g^*(\cdot \mid L)$:

  \[
  \mathbb{E} Y_{g^*} = \int_L \int_s \mathbb{E}(Y \mid S = s, L)g(s - \delta \mid L) \cdot ds \cdot dP_0(L)
  \]

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

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

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

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

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

  \[
  = \beta \delta
  \]
References


