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

Nima Hejazi

Thursday, 15 October 2020

Graduate Group in Biostatistics, and Center for Computational Biology, University of California, Berkeley

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

Biomedical Big Data Seminar, UC Berkeley, Fall 2020
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?
HITN 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 immunogenic response (due to vaccine) differed?

- Immunogenic response profiles only available for second-stage 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 $X = (L, A, Y) \sim P^X_0 \in \mathcal{M}^X$, as per the full HVTN 505 trial cohort (Hammer et al. 2013):
  - $L$ (baseline covariates): sex, age, BMI, behavioral HIV risk,
  - $A$ (exposure): immune response profile for CD4+ and CD8+,
  - $Y$ (outcome of interest): HIV-1 infection status at week 28.
- Observed data $O = (C, CX) = (L, C, CA, Y)$; $C \in \{0, 1\}$ is an indicator for inclusion in the second-stage sample.
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 \textit{a priori}. 
**Stochastic interventions** modify the value $A$ would naturally assume by drawing from a modified exposure distribution.

Consider the post-intervention value $A^* \sim G^*(\cdot | L)$; 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 \( \Psi(P_0^X) = \mathbb{E}_{P_0^{X,L}} \overline{Q}(d(A, L), L) \), 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 \( G^*(\cdot \mid L) \).
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:

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

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.
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 $(g_0(L), \overline{Q}_0(A, L)) \times (\pi_0(Y, L), \mathbb{E}(D^F(P_0^X)(x) \mid C = 1, Y, L)).$
Identifying the best efficient estimator

Figure 1: Relative performance of reweighted and augmented estimators.
Fighting the HIV-1 epidemic with preventive vaccines

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

But what about \( O = (L, A, Z, CM, \Delta, \tilde{T}, C) \)?
- \( \tilde{T} = min(T, C) \): possibly right-censored time to symptomatic SARS-CoV-2 infection
- \( \Delta = I(T < C) \): observed 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 \( \{M, Y\} \) relationship.
  - A3: No unmeasured confounding of \( \{A, M\} \) relationship.
  - A4: No \( \{M, Y\} \) 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 \( M \), no \( Z \), “cross-world” independence.
  - Stochastic (in)direct effects (Díaz and Hejazi 2020): continuous \( A \) and \( M \), no \( Z \); no “cross-world” exclusion.
  - Interventional (in)direct effects (Díaz et al. 2020): binary \( A \), continuous \( M \), \( Z \) ok, no “cross-world” exclusion.
  - Stochastic interventional (in)direct effects (Hejazi et al. 2020): continuous \( A \) and \( M \), \( Z \) ok, no “cross-world” exclusion.
Appendix
From the causal to the statistical target parameter

**Assumption 1: Consistency**

\[ Y_{i}^{d(a_{i}, l_{i})} = Y_{i} \text{ in the event } A_{i} = d(a_{i}, l_{i}), \text{ for } i = 1, \ldots, n \]

**Assumption 2: SUTVA**

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

**Assumption 3: Strong ignorability**

\[ A_{i} \perp Y_{i}^{d(a_{i}, l_{i})} \mid L_{i}, \text{ for } i = 1, \ldots, n \]
Assumption 4: *Positivity (or overlap)*

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

- 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 \( a_i \in \mathcal{A} \) and \( l_i \in \mathcal{L} \).
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:

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

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.
Nonparametric conditional density estimation

- 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

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

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 \( A_i \) belongs, and the binary variables indicating \( A_i \in [\alpha_{t-1}, \alpha_t) \) are recorded.
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.
Key properties of TML estimators

- **Asymptotic linearity:**

  \[
  \Psi(P^*_n) - \Psi(P^{X}_0) = \frac{1}{n} \sum_{i=1}^{n} D(P^{X}_0)(X_i) + o_P \left( \frac{1}{\sqrt{n}} \right)
  \]

- **Gaussian limiting distribution:**

  \[
  \sqrt{n}(\Psi(P^*_n) - \Psi(P^{X}_0)) \rightarrow N(0, \text{Var}(D(P^{X}_0)(X)))
  \]

- **Statistical inference:**

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

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

1. Construct initial estimators $g_n$ of $g_0(A, L)$ 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 $\overline{Q}_n^{\star}$ of the initial estimate $\overline{Q}_{n, \epsilon_n}$:

   \[ \Psi_n = \Psi(P_n^{\star}) = \frac{1}{n} \sum_{i=1}^{n} \overline{Q}_n^{\star}(d(A_i, L_i), L_i). \]
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 second-stage sample \( C = 1 \), construct initial estimators \( g_n(A, L) \) and \( \overline{Q}_n(A, L) \), 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 \( \frac{1}{n} \sum_{i=1}^{n} \text{EIF}_i < \frac{1}{n} \).
A linear modeling perspective

- 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

\[
EY_{A+\delta} - EY_A = [\beta_0 + \beta_1 (EA + \delta)] - [\beta_0 + \beta_1 (EA)]
\]
\[
= \beta_0 - \beta_0 + \beta_1 EA - \beta_1 EA + \beta_1 \delta
\]
\[
= \beta_1 \delta
\]

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

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

- Thus, a unit shift for \(a' \in \mathcal{A}\) (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, $\mathbb{E}(Y \mid A = z, L) = \beta z + \theta(L)$:

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

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

$$= \beta \delta$$


