

# Vaccine efficacy assessment under two-phase sampling based on the causal effects of stochastic interventions

---


Nima Hejazi

Wednesday, 05 June 2019

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

 nshejazi

 nhejazi

 nimahejazi.org

 [bit.ly/2019\\_sfasa\\_jsm](https://bit.ly/2019_sfasa_jsm)

joint work with David Benkeser and Mark van der Laan



# 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.
- **Open question:** How can HIV-1 vaccines be improved by modulating immunogenic CD4+ or CD8+ response profiles?

## HVTN 505 trial examined new antibody boost vaccines

- HIV Vaccine Trials Network (HVTN) 505 vaccine efficacy RCT with  $n = 2504$  (Hammer et al. 2013) participants.
- In vaccination arm, immunogenic response profiles only made available for second-stage sample  $n = 189$  (Janes et al. 2017).
- Two-phased sampling mechanism: 100% inclusion rate if HIV-1 positive in week 28; variable rate otherwise.
- **Question:** How would HIV-1 infection risk in week 28 have differed had immunogenic response (due to vaccine) differed?

## Two-phase sampling censors the complete data structure

- Complete, unobserved data  $X = (W, A, Y) \sim P_0^X \in \mathcal{M}_{NP}^X$ , as per the full HVTN 505 RCT (Hammer et al. 2013):
  - $W$  — baseline covariates: sex, age, BMI, behavioral HIV risk,
  - $A$  — intervention: immune response profile for CD4 and CD8,
  - $Y$  — outcome of interest: HIV-1 infection status as of week 28.
- Observed data  $O = (\Delta, \Delta X) = (W, \Delta, \Delta A, Y)$ ,  $\Delta \in \{0, 1\}$ , as per the second-stage sample of Janes et al. (2017).

## NPSEM for the (uncensored) full data $X$

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

$$W = f_W(U_W)$$

$$A = f_A(W, U_A)$$

$$Y = f_Y(A, W, U_Y)$$

- NPSEM parameterizes likelihood  $p_0^X$  in terms of the distribution of RVs  $(X, U)$  modeled by this system.
- Implies a model for the distribution of counterfactual RVs generated by interventions on the data-generating process.

# Stochastic interventions alter the NPSEM

- *Stochastic interventions* modify the value  $A$  would naturally assume by replacing  $f_A(W, U_A)$ .
- How? By drawing from a modified intervention distribution  $G^*(\cdot \mid W)$ , i.e.,  $A^* \sim G^*(\cdot \mid W)$ .
- This generates a counterfactual RV, with distribution  $P_0^d$ ,  $Y_{G^*} := f_Y(A^*, W, U_Y)$ .
- We estimate  $\psi_{0,d} := \mathbb{E}_{P_0^d}\{Y_{d(A,W)}\}$ , mean of  $Y_{d(A,W)}$ , where the rule  $d(A, W)$  defines  $G^*(\cdot \mid W)$ .

## Literature: Díaz and van der Laan (2012)

- Identification conditions for a statistical parameter of the counterfactual outcome  $\psi_{0,d}$  under such interventions.
- Show that the causal quantity of interest  $\mathbb{E}_{P_0^d}\{Y_{d(A,W)}\}$  is identified by a functional of the distribution of  $X$ :

$$\psi_{0,d} = \int_{\mathcal{W}} \int_{\mathcal{A}} \mathbb{E}_{P_0^X}\{Y \mid A = d(a, w), W = w\} \cdot q_{0,A}^X(a \mid W = w) \cdot q_{0,W}^X(w) d\mu(a) d\nu(w)$$

- Provides a derivation based on the efficient influence function (EIF) with respect to the nonparametric model  $\mathcal{M}$ .

# Identifying the causal parameter from the observed data

## Assumption 1: *Consistency*

$Y_i^{d(a_i, w_i)} = Y_i$  in the event  $A_i = d(a_i, w_i)$ , for  $i = 1, \dots, n$

## Assumption 2: *SUTVA*

$Y_i^{d(a_i, w_i)}$  does not depend on  $d(a_j, w_j)$  for  $i = 1, \dots, n$  and  $j \neq i$ , or lack of interference (Rubin 1978; 1980)

## Assumption 3: *Strong ignorability*

$A_i \perp\!\!\!\perp Y_i^{d(a_i, w_i)} \mid W_i$ , for  $i = 1, \dots, n$



### **Assumption 4: *Positivity (or overlap)***

$a_i \in \mathcal{A} \implies d(a_i, w_i) \in \mathcal{A}$  for all  $w \in \mathcal{W}$ , where  $\mathcal{A}$  denotes the support of  $A$  conditional on  $W = w_i$  for all  $i = 1, \dots, n$

- Does not require the intervention density place mass across all strata defined by  $W$ .
- Rather, merely requires the post-intervention quantity be seen in the observed data for given  $a_i \in \mathcal{A}$  and  $w_i \in \mathcal{W}$ .

# 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?

$$d(a, w) = \begin{cases} a + \delta, & \text{if plausible} \\ a, & \text{otherwise} \end{cases}$$

- Díaz and van der Laan (2012; 2018) give a statistical target parameter and influence function for the complete data case:

$$\psi(P_0^X) = \mathbb{E}_{P_0^X} \bar{Q}(d(A, W), W),$$

allowing estimation of causal parameter  $\psi_{0,d} = \mathbb{E} Y_{d(A,W)}$ .

# HIV-1 risk under stochastically shifted immune responses

## Targeted minimum loss estimation (TMLE)

- A TMLE algorithm updates initial estimators (e.g., via logistic tilting) so as to satisfy a set of estimating equations.
- Semiparametric-efficient estimation thru solving efficient influence function estimating equation wrt the model  $\mathcal{M}$ .
- For  $\Psi(P_0^X)$  which the efficient influence function (EIF) is

$$D(P_0^X)(x) = H(a, w)(y - \bar{Q}(a, w)) + \bar{Q}(d(a, w), w) - \Psi(P_0^X)$$

- The auxiliary covariate  $H(a, w)$  may be expressed

$$H(a, w) = \mathbb{I}(a < u(w)) \frac{g_0(a - \delta \mid w)}{g_0(a \mid w)} + \mathbb{I}(a \geq u(w) - \delta)$$

## Consistent estimation in spite of two-phase sampling

- What if sampling mechanism  $\pi_0(Y, W) = \mathbb{P}(\Delta = 1 \mid Y, W)$  is not known by design? Nonparametric estimation of  $\pi_0(Y, W)$ ?
- Building on Rose and van der Laan (2011), we provide
  - asymptotically linear and nonparametric-*efficient* estimators;
  - multiply *robust*, with 2 forms of double robustness;
  - Gaussian limiting distributions and Wald-type CIs.
- *Initial proposal*: Use an IPC-weighted loss function

$$\mathcal{L}(P_0^X)(O) = \frac{\Delta}{\pi_n(Y, W)} \mathcal{L}^F(P_0^X)(X)$$

## Efficient estimation in spite of two-phase sampling

- When  $\pi_0(Y, W)$  is estimated nonparametrically, adding IPC weights to the loss is insufficient for estimator efficiency.
- Instead, an EIF augmented with IPC weights must be used

$$D(P_0^X)(o) = \frac{\Delta}{\pi_0(y, w)} D^F(P_0^X)(x) \\ - \left(1 - \frac{\Delta}{\pi_0(y, w)}\right) \mathbb{E}(D^F(P_0^X)(x) \mid \Delta = 1, Y = y, W = w),$$

expressed in terms of the full data EIF  $D^F(P_0^X)(x)$ .

## Efficient estimation in spite of two-phase sampling

The IPC-augmented EIF has two distinct terms

$$\frac{\Delta}{\pi_0(y,w)} D^F(P_0^X)(x)$$

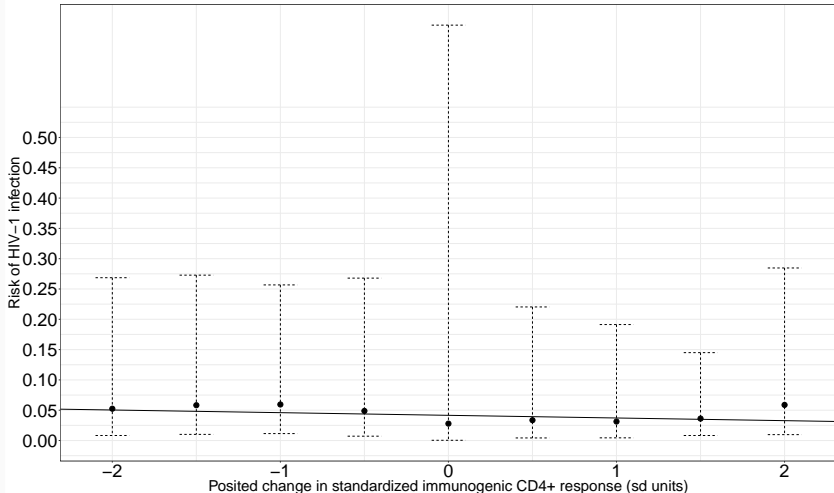
The IPC-weighted EIF of full data structure  $X$  relative to  $\mathcal{M}$ ; and,

$$\left(1 - \frac{\Delta}{\pi_0(y,w)}\right) \mathbb{E}(D^F(P_0^X)(x) \mid \Delta = 1, Y = y, W = w)$$

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

# How does this help in fighting the HIV-1 epidemic? CD4+

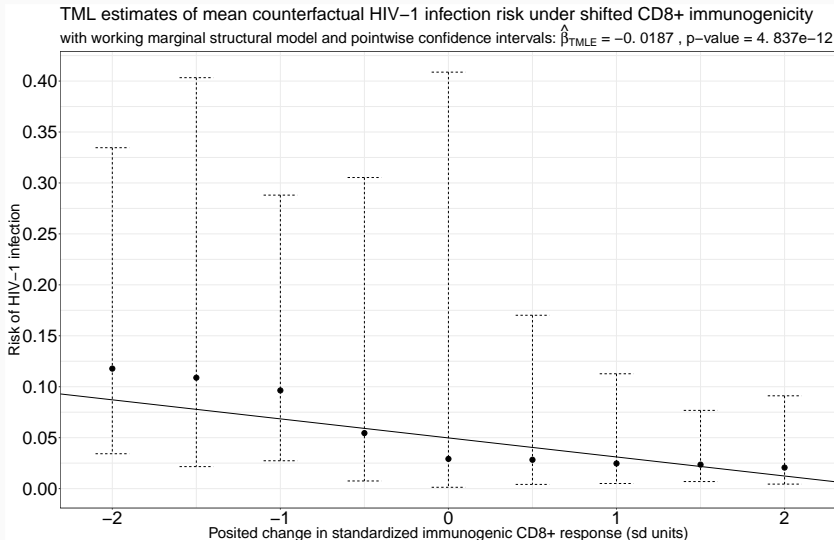
TML estimates of mean counterfactual HIV-1 infection risk under shifted CD4+ immunogenicity with working marginal structural model and pointwise confidence intervals:  $\hat{\beta}_{\text{TMLE}} = -0.0044$ ,  $p\text{-value} = 1.798\text{e-}01$



**Figure 1:** Analysis of HIV-1 risk as a function of CD4+ immunogenicity, using R package txshift (<https://github.com/nhejazi/txshift>.)



# How does this help in fighting the HIV-1 epidemic? CD8+



**Figure 2:** Analysis of HIV-1 risk as a function of CD8+ immunogenicity, using R package txshift (<https://github.com/nhejazi/txshift>.)

# Efficient and robust estimation under two-phase sampling

- We now have a semiparametric-efficient and robust procedure for assessing the effect of the intervention  $d(a, w) = a + \delta$ .
- Due to construction based on the IPCW-EIF, any resultant estimators are robust and efficient under two-phase sampling.
- New *causal* tool for assessing how immunogenic response shifts would have affected HIV-1 infection risk.
- New open source software for deploying such estimators:
  - <https://github.com/nhejazi/haldensify> (densities)
  - <https://github.com/nhejazi/txshift> (AIPW, TMLE)
  - <https://github.com/tlverse/tmle3shift> (TMLE)

# References

---


- Díaz, I. and van der Laan, M. J. (2011). Super learner based conditional density estimation with application to marginal structural models. *The international journal of biostatistics*, 7(1):1–20.
- Díaz, I. and van der Laan, M. J. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549.
- Díaz, I. and van der Laan, M. J. (2018). Stochastic treatment regimes. In *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pages 167–180. Springer Science & Business Media.
- Hammer, S. M., Sobieszczyk, M. E., Janes, H., Karuna, S. T., Mulligan, M. J., Grove, D., Koblin, B. A., Buchbinder, S. P., Keefer, M. C., Tomaras, G. D., et al. (2013). Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. *New England Journal of Medicine*, 369(22):2083–2092.


- Janes, H. E., Cohen, K. W., Frahm, N., De Rosa, S. C., Sanchez, B., Hural, J., Magaret, C. A., Karuna, S., Bentley, C., Gottardo, R., et al. (2017). Higher t-cell responses induced by DNA/rAd5 HIV-1 preventive vaccine are associated with lower HIV-1 infection risk in an efficacy trial. *The Journal of infectious diseases*, 215(9):1376–1385.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- Rose, S. and van der Laan, M. J. (2011). A targeted maximum likelihood estimator for two-stage designs. *The International Journal of Biostatistics*, 7(1):1–21.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.
- Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.
- van der Laan, M. J., Dudoit, S., and Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–23.
- van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

# Thank you.

Slides: [bit.ly/2019\\_sfasa\\_jsm](https://bit.ly/2019_sfasa_jsm)



 <https://nimahejazi.org>

 <https://github.com/nhejazi>

 <https://twitter.com/nshejazi>

# Appendix

# Nonparametric Conditional Density Estimation

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

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

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.

# Nonparametric Conditional 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

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

- 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  $g(A \mid W)$  and  $g(A - \delta \mid W)$ , 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.

## Algorithm for IPCW-TML Estimation

1. Using all observed units ( $X$ ), estimate sampling mechanism  $\pi(Y, W)$ , perhaps using data-adaptive regression methods.
2. Using only observed units in the second-stage sample  $\Delta = 1$ , construct initial estimators  $g_n(A, W)$  and  $\bar{Q}_n(A, W)$ , weighting by the sampling mechanism estimate  $\pi_n(Y, W)$ .
3. With the approach described for the full data case, compute  $H_n(a_i, w_i)$ , 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  $\pi_n(Y, W)$  and IPCW-augmented EIF, updating TML estimate in each iteration, until  $\frac{1}{n} \sum_{i=1}^n \text{EIF}_i < \frac{1}{n}$ .