# 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.

- Baseline covariates($L$): sex, age, BMI, behavioral HIV risk.
- Intervention(s) ($A$): post-vaccination T-cell activity markers.
- Outcome ($Y$): HIV-1 infection status at week 28 of tiral.
- 12-color intracellular cytokine staining (ICS) assay.
- Cryopreserved peripheral blood mononuclear cells were stimulated with synthetic HIV-1 peptide pools.
- All immune responses are assayed *after* the endpoints of interest (HIV-1 infection status) are collected.
- **Conclusion:** Understanding which immune responses impact vaccine efficacy helps develop more efficacious vaccines.
- A vaccine effective at preventing HIV-1 acquisition would be a cost-effective and durable approach to halting the worldwide epidemic.

## 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., $O = X \mid A = 1$.

- $P_0^X$ — true (unknown) distribution of the full data $X$,
- $\mathcal{M}_{NP}^X$ — nonparametric statistical model.

## 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*.

## NPSEM with stochastic interventions

- *Stochastic interventions* modify the value $S$ would naturally assume by drawing from a modified exposure distribution.
- Consider the post-intervention value $S^\star \sim G^\star(\cdot \mid L)$; static interventions are a special case (degenerate distribution).
- Such an intervention generates a counterfactual random variable $Y_{G^\star} := f_Y(S^\star, L, U_Y)$, with distribution $P_0^\delta$, .
- We aim to estimate $\psi_{0,\delta} := \mathbb{E}_{P_0^\delta}\{Y_{G^\star}\}$, the counterfactual mean under the post-intervention exposure distribution $G^\star$.

## Stochastic interventions for the causal effects of shifts

- 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 $\psi_{0,d} := \mathbb{E}_{P_0^d}\{Y_{d(S,L)}\}$, mean of $Y_{d(S,L)}$.
- Statistical target parameter is $\Psi(P_0^X) = \mathbb{E}_{P_0^X}\overline{Q}(d(S, L), L)$, 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 $G^{\star}(\cdot \mid L)$.

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

## From the causal to the statistical target parameter

**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***

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

**Assumption 3: *Positivity***

$s_i \in \mathcal{S} \implies d(s_i, l_i) \in \mathcal{S}$ *for all* $l \in \mathcal{L}$*, where* $\mathcal{S}$ *denotes the support of* $S$ *conditional on* $L = l_i$ *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 $s_i \in \mathcal{S}$ and $l_i \in \mathcal{L}$.

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

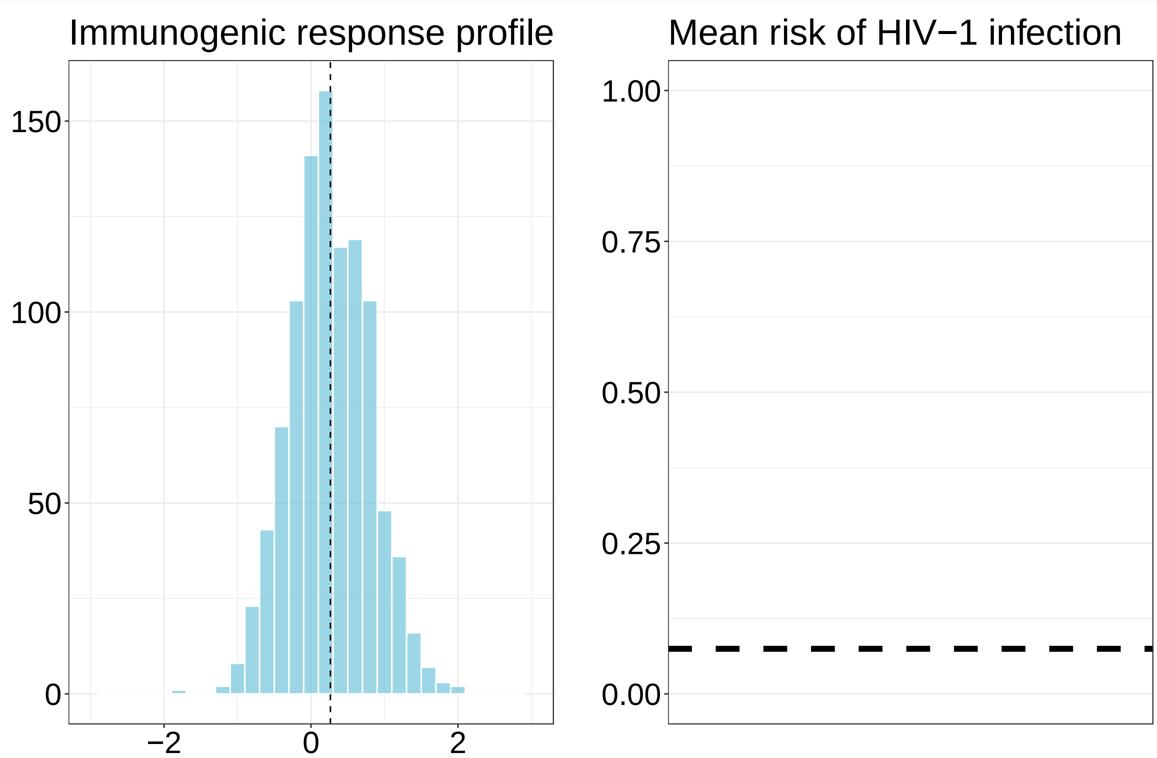
- *Proposal:* Evaluate outcome under an altered *intervention distribution* — e.g., $P_{\delta}(g_0)(S = s \mid L) = g_0(s - \delta(L) \mid L)$.
- Identification conditions for a statistical parameter of the counterfactual outcome $\psi_{0,d}$ under such an intervention.
- Show that the causal quantity of interest $\mathbb{E}_0\{Y_{d(S,L)}\}$ is identified by a functional of the distribution of $X$:

$$\psi_{0,d} = \int_{\mathcal{L}} \int_{\mathcal{S}} \mathbb{E}_{P_0^X}\{Y \mid S = d(s,l), L = l\} \cdot q_{0,S}^X(s \mid L = l) \cdot q_{0,L}^X(l) d\mu(s) d\nu(l)$$

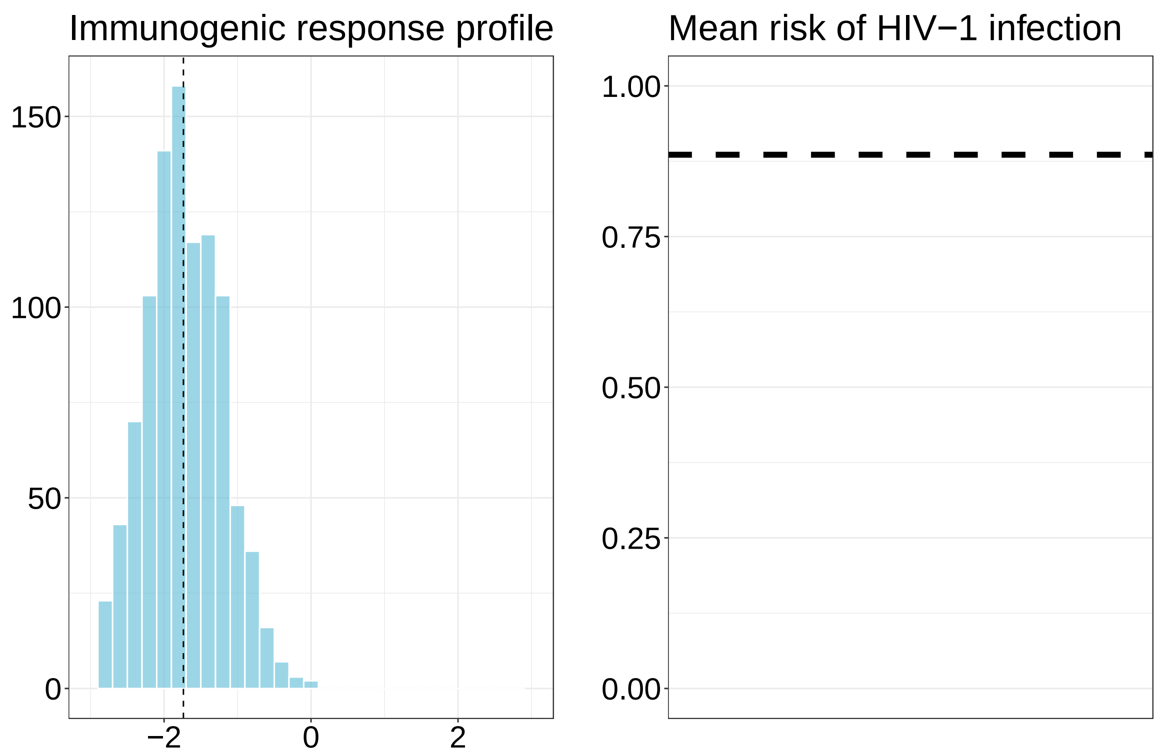
- Provides a derivation based on the efficient influence function (EIF) with respect to the nonparametric model $\mathcal{M}$.

- The identification result allows us to write down the causal quantity of interest in terms of a functional of the observed data.
- Key innovation: loosening standard assumptions through a change in the observed intervention mechanism.
- Problem: globally altering an intervention mechanism does not necessarily respect individual characteristics.
- The authors build IPW, one-step, and TML estimators, comparing the three different approaches.

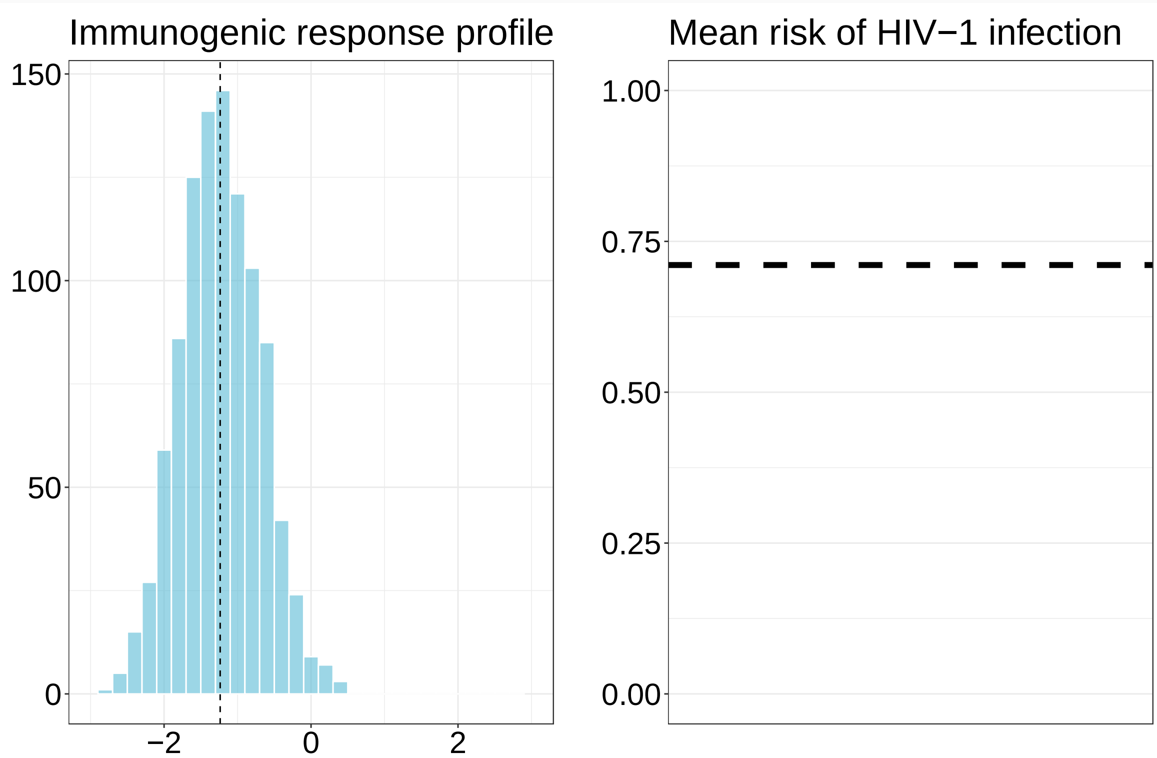
## HIV-1 risk under shifted immunogenic responses



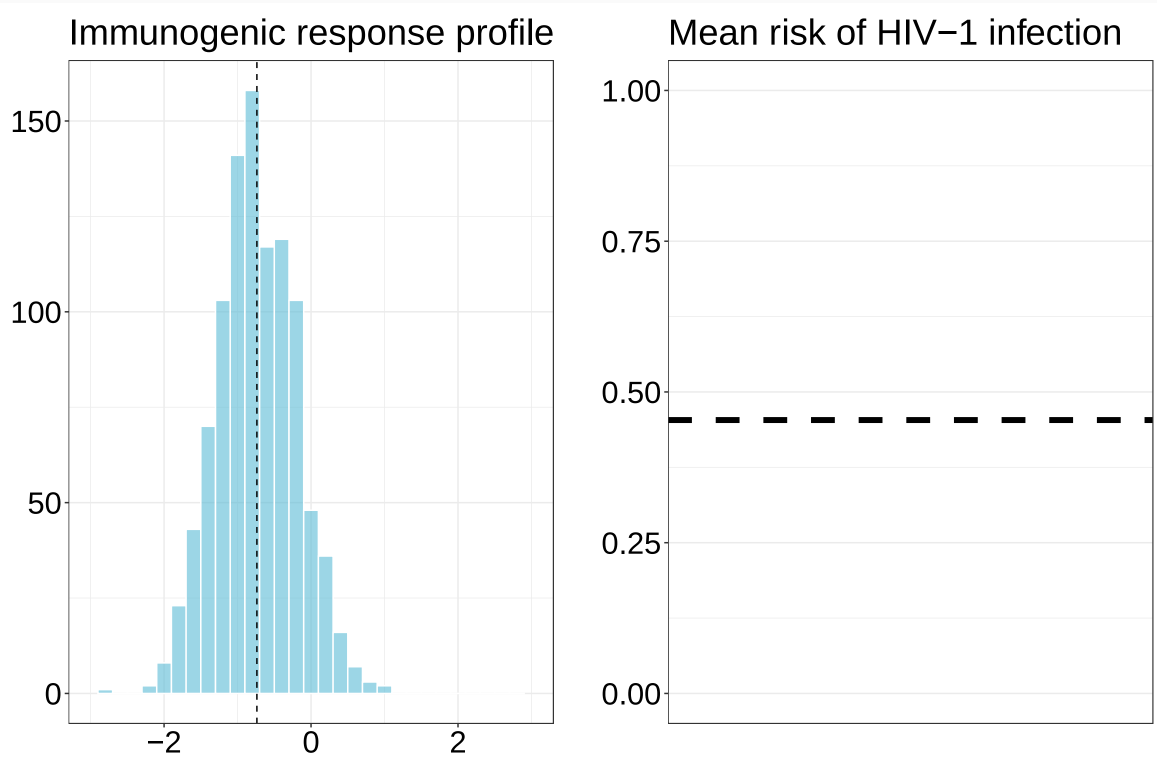
## HIV-1 risk under shifted immunogenic responses



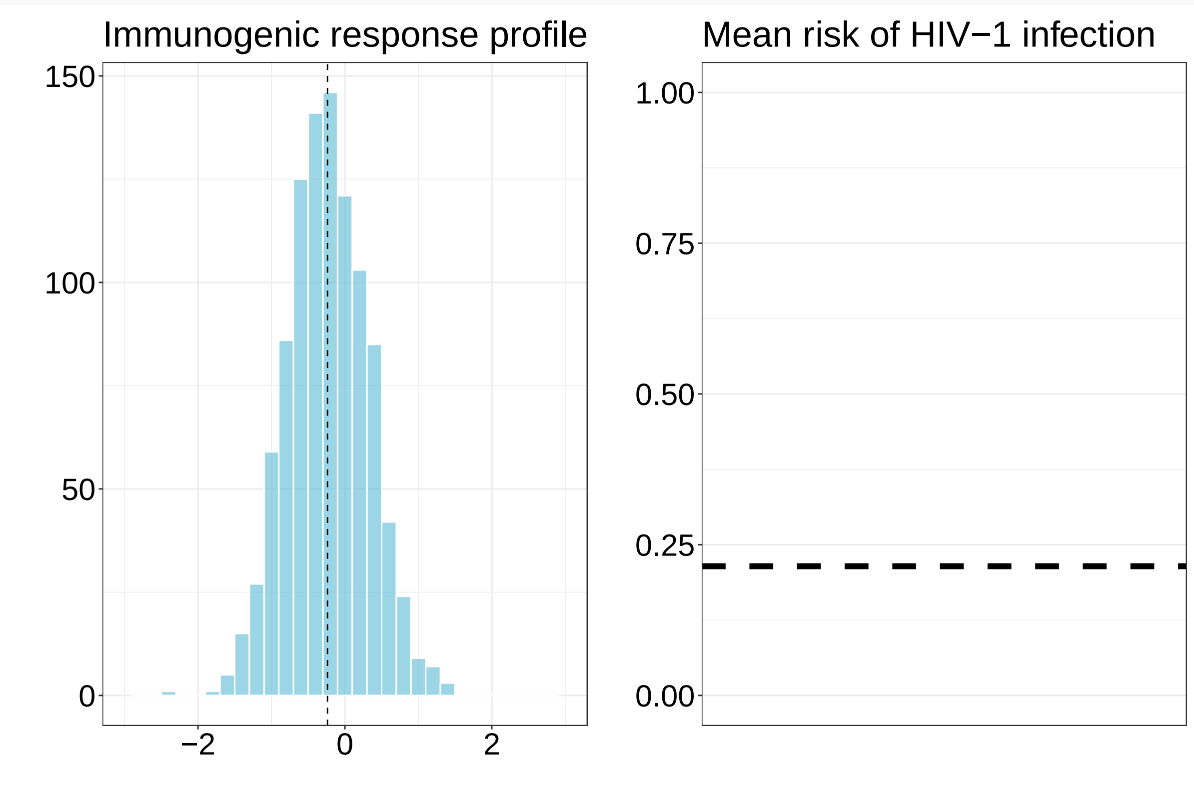
## HIV-1 risk under shifted immunogenic responses



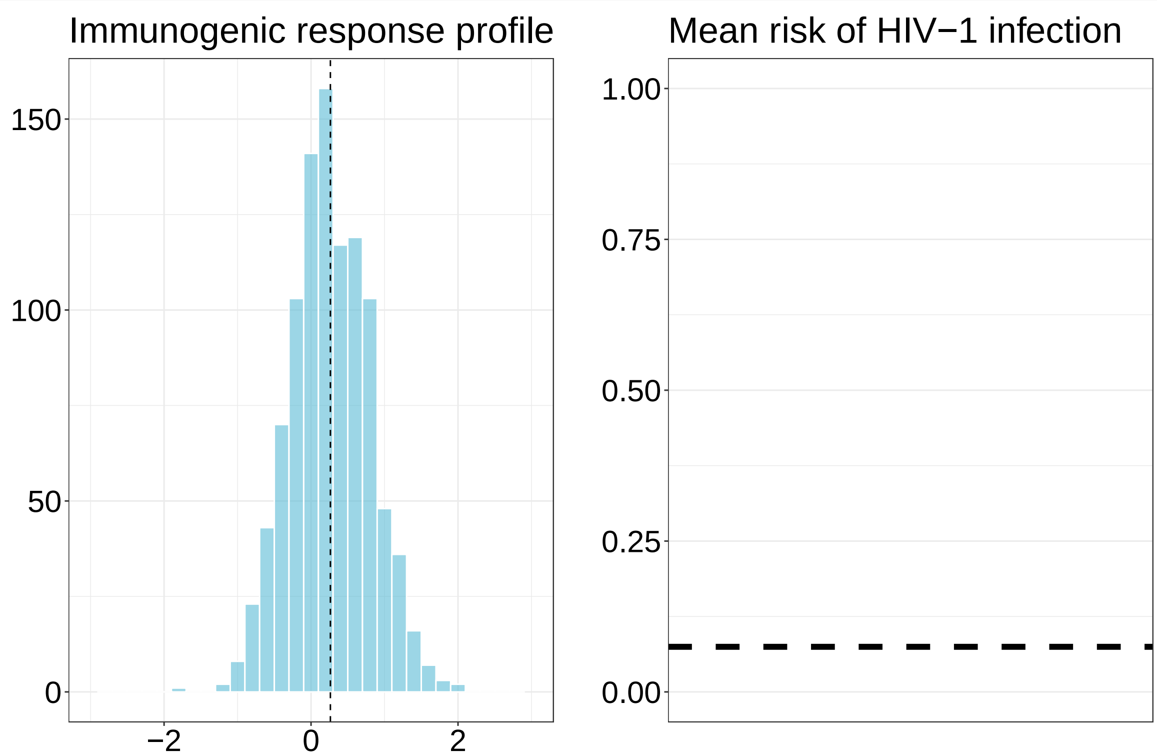
## HIV-1 risk under shifted immunogenic responses



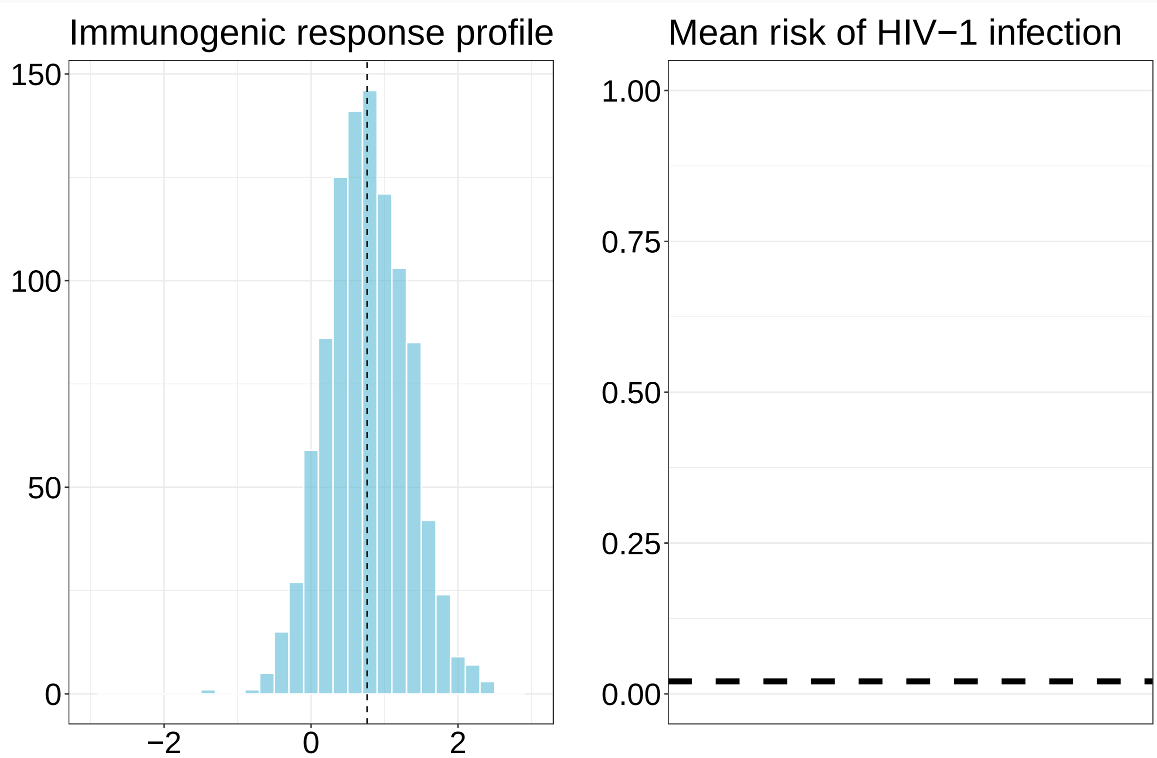
## HIV-1 risk under shifted immunogenic responses



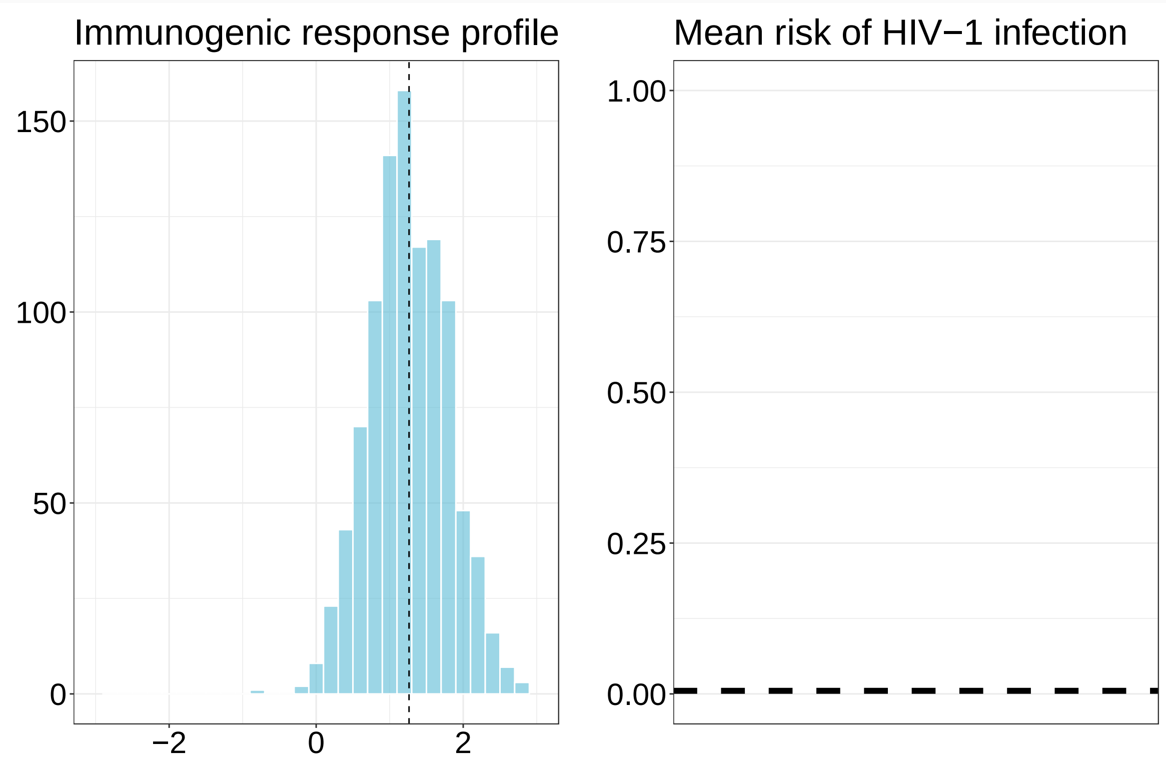
## HIV-1 risk under shifted immunogenic responses



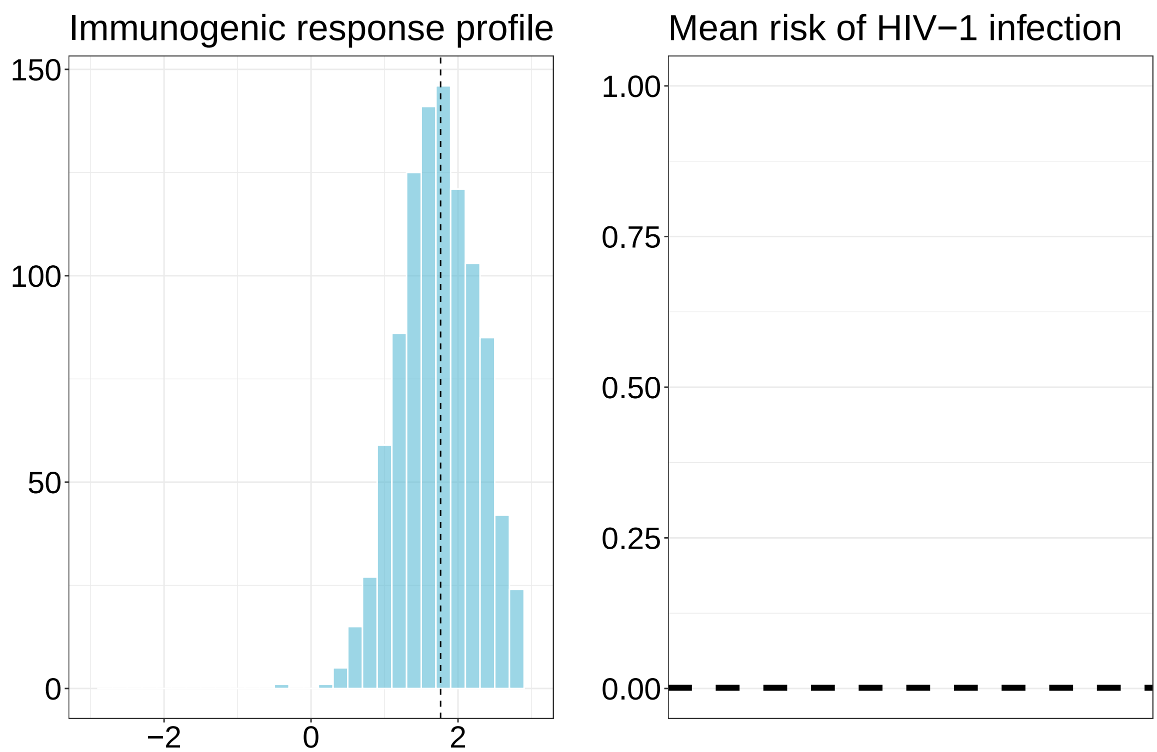
## HIV-1 risk under shifted immunogenic responses



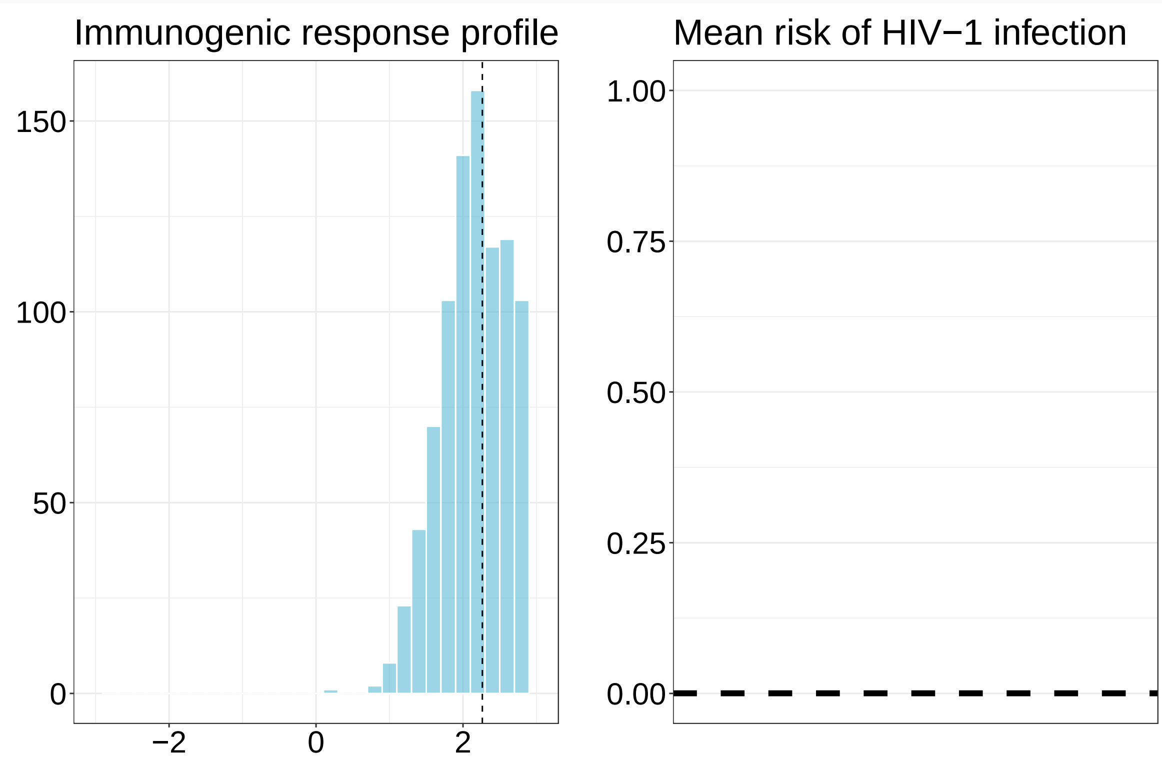
## HIV-1 risk under shifted immunogenic responses



## HIV-1 risk under shifted immunogenic responses



## HIV-1 risk under shifted immunogenic responses



## Flexible, efficient estimation

- The efficient influence function (EIF) is:

$$D(P_0^X)(x) = H(s,l)(y - \overline{Q}(s,l)) + \overline{Q}(d(s,l),l) - \Psi(P_0^X).$$

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

$$\Psi_n^+ = \frac{1}{n}\sum_{i=1}^{n} \overline{Q}_n(d(S_i,L_i),L_i) + D_n(O_i).$$

- The TML estimator updates initial estimates of $\overline{Q}_n$ by tilting:

$$\Psi_n^\star = \frac{1}{n}\sum_{i=1}^{n} \overline{Q}_n^\star(d(S_i,L_i),L_i).$$

- Both estimators are doubly robust.

- Both estimators are CAN even when nuisance parameters are estimated via flexible, machine learning techniques.
- Semiparametric-efficient estimation thru solving efficient influence function estimating equation wrt the model $\mathcal{M}$.
- The auxiliary covariate simplifies when the treatment is in the limits (conditional on $W$) — i.e., for $S_i \in (u(l) - \delta, u(l))$, then we have $H(s, l) = \frac{g_0(s-\delta|l)}{g_0(s|l)} + 1$.
- Need to explicitly remind the audience what $u(l)$ is again. It's only appeared once at this point, and only been mentioned in passing.

## Augmented estimators for two-phase sampling designs

- 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

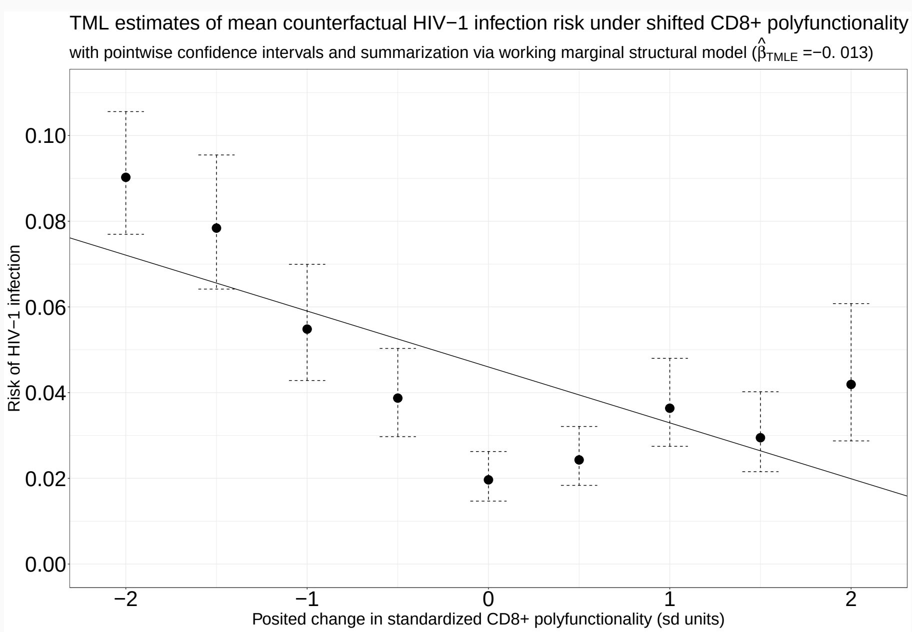



**Figure 1:** Analysis of HIV-1 risk as a function of CD8+ immunogenicity, using `R` package `txshift` (`https://github.com/nhejazi/txshift`).

## 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

- Nucleic acid vaccines have never been approved before, but are quick to manufacture.
- Viral-vectored vaccines are also quick to manufacture but can develop immunity against vector.
- Subunit vaccines are a construct of several effective vaccines, but are slower to manufacture and often require an adjuvant.

## 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.

## Immune Correlates of Protection (Plotkin and Gilbert 2012)

- 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.

## Measuring Correlates: Two-Phase Designs

- 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 ($\approx$ 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:

$$\text{SVE}(\delta) = 1 - \frac{\mathbb{E}[\mathbb{P}(Y = 1 \mid A = 1, S = s + \delta, L = l) \mid A = 1, L]}{\mathbb{P}(Y(0) = 1)}.$$

- $\mathbb{P}(Y(0) = 1)$: 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 $\delta$.
- Further details in CoVPN's public immune correlates SAP at `https://doi.org/10.6084/m9.figshare.13198595`.

## 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 $O = (L, A, Z, CS, \Delta, \widetilde{T}, C)$?
  - $\widetilde{T} = \min(T_F, T_C)$: 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 $\widetilde{T}$?

- Goal: assess *indirect* effect of vaccination through mCoPs.
- Define/identify new mCoPs to be used as surrogate endpoints.
- Could also have missing outcome in the binary endpoint case.

## 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.

- A1, A3 hold in randomized trials.
- A2 may not hold: include all mutual $\{S, Y\}$ predictors, then perform sensitivity analysis.
- A4 usually doesn't hold: either measure $S$ right after $A$ or develop more flexible effect definitions.
- "Cross-world" independence: $Y(a, s) \perp S(a') \quad \forall s$; untestable in RCTs
- Extensions for two-phase sampling?

# Appendix

## Literature: Haneuse and Rotnitzky (2013)

- *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).

---

- Shifts of the form $d(S, L)$ are considerably more interesting since these are realistic intervention policies.
- Example: consider an individual with an extremely high immune response but whose baseline covariates $L$ suggest we shift the response still higher. Such a shift may not be biologically plausible (impossible, even) but we cannot account for this if the shift is only a function of $L$.
- The authors build IPW, outcome regression, and non-iterative doubly robust estimators, as well as an approach based on MSMs.
- Piecewise smooth invertibility: This assumption ensures that we can use the change of variable formula when computing integrals over $S$ and it is useful to study the estimators that we propose in this paper.

## Literature: Young et al. (2014)

- 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 $\mathbb{E}Y_{d(S,L)}$ or $\mathbb{E}Y_{d(L)}$.
- The authors also consider limits on implementing shifts $d(S, L)$, and address working in a longitudinal setting.

## Literature: Díaz and van der Laan (2018)

- Builds on the original proposal, accomodating 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.

## Nonparametric conditional 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.

## Density estimation with the Super Learner algorithm

- 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.

---

- The auxiliary covariate simplifies when the treatment is in the limits (conditional on $L$) — i.e., for $S_i \in (u(l) - \delta, u(l))$, then we have $H(s, l) = \frac{g_0(s-\delta|l)}{g_0(s|l)} + 1$.
- Asymptotically optimal in the sense that it performs as well as the oracle selector as the sample size increases.

## Key properties of TML estimators

- **Asymptotic linearity:**

$$\Psi(P_n^\star) - \Psi(P_0^X) = \frac{1}{n}\sum_{i=1}^{n} D(P_0^X)(X_i) + o_P\left(\frac{1}{\sqrt{n}}\right)$$

- **Gaussian limiting distribution:**

$$\sqrt{n}(\Psi(P_n^\star) - \Psi(P_0^X)) \to N(0, Var(D(P_0^X)(X)))$$

- **Statistical inference:**

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

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

Under the additional condition that the remainder term $R(\hat{P}^\star, P_0)$ decays as $o_P\left(\frac{1}{\sqrt{n}}\right)$, we have that $\Psi_n - \Psi_0 = (P_n - P_0)\cdot D(P_0) + o_P\left(\frac{1}{\sqrt{n}}\right)$, which, by a central limit theorem, establishes a Gaussian limiting distribution for the estimator, with variance $V(D(P_0))$, the variance of the efficient influence function when $\Psi$ admits an asymptotically linear representation.
The above implies that $\Psi_n$ is a $\sqrt{n}$-consistent estimator of $\Psi$, that it is asymptotically normal (as given above), and that it is locally efficient. This allows us to build Wald-type confidence intervals, where $\sigma_n^2$ is an estimator of $V(D(P_0))$. The estimator $\sigma_n^2$ may be obtained using the bootstrap or computed directly via $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} D^2(\bar{Q}_n^\star, g_n)(O_i)$
We obtain semiparametric-efficient estimation and robust inference in the nonparametric model $\mathcal{M}$ by solving the efficient influence function.

1. If $D(\bar{Q}_n^\star, g_n)$ converges to $D(P_0)$ in $L_2(P_0)$ norm.
2. The size of the class of functions $\bar{Q}_n^\star$ and $g_n$ is bounded (technically, $\exists \mathcal{F}$ s.t. $D(\bar{Q}_n^\star, g_n) \in \mathcal{F}$ w.h.p., where $\mathcal{F}$ is a Donsker class)

## Algorithm for TML estimation

1. Construct initial estimators $g_n$ of $g_0(S, L)$ and $Q_n$ of $\overline{Q}_0(S, L)$, 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
$$\text{logit}\overline{Q}_{\epsilon,n}(s, l) = \text{logit}\overline{Q}_n(s, l) + \epsilon H_n(s, l),$$
or an alternative regression model incorporating weights.
4. Compute TML estimator $\Psi_n$ of the target parameter, defining update $\overline{Q}_n^{\star}$ of the initial estimate $\overline{Q}_{n,\epsilon_n}$:
$$\Psi_n = \Psi(P_n^{\star}) = \frac{1}{n}\sum_{i=1}^{n} \overline{Q}_n^{\star}(d(S_i, L_i), L_i).$$

- We recommend using nonparametric methods for the initial estimators, as consistent estimation is necessary for efficiency of the estimator $\Psi_n$.
- Intuition for the submodel fluctuation?

## 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 $\overline{Q}_n(S, L)$, 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 $\frac{1}{n}\sum_{i=1}^{n} \text{EIF}_i < \frac{1}{n}$.

- We recommend using nonparametric methods for the initial estimators, as consistent estimation is necessary for efficiency of the estimator $\Psi_n$.
- Intuition for the submodel fluctuation?
- This process includes the use of HAL to fit the regression of the EIF contributions on the sampling node $\{Y, L\}$.

# Identifying the best efficient estimator

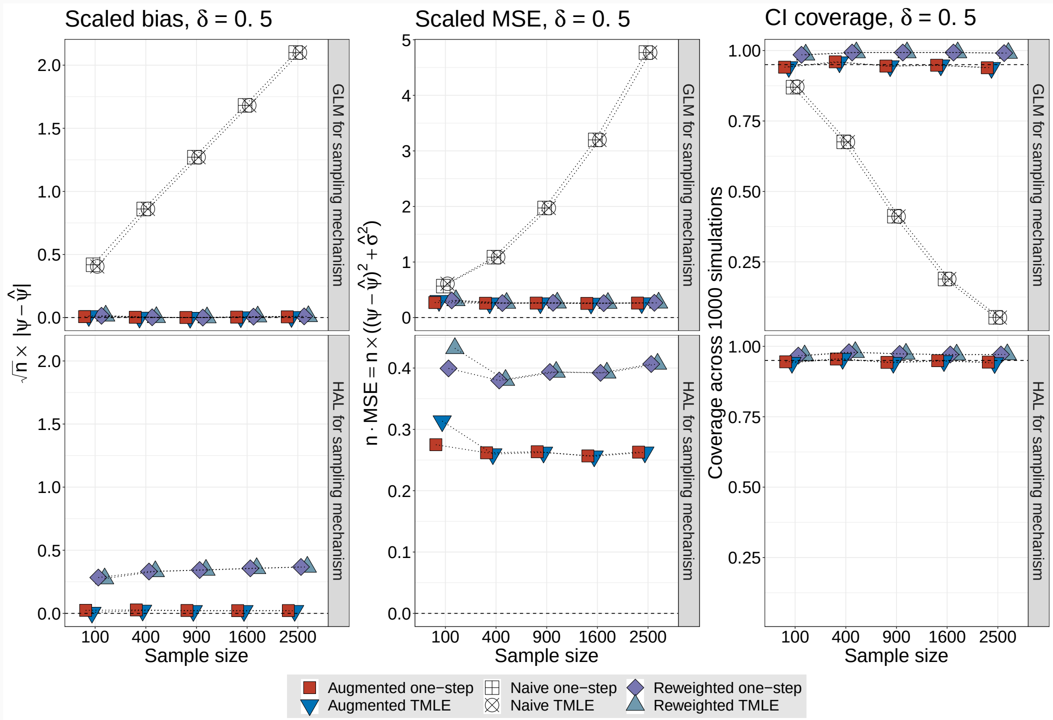



**Figure 2:** Relative performance of reweighted and augmented estimators.

## 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

$$
\begin{aligned}
\mathbb{E}Y_{S+\delta} - \mathbb{E}Y_S &= [\beta_0 + \beta_1(\mathbb{E}S + \delta)] - [\beta_0 + \beta_1(\mathbb{E}S)] \\
&= \beta_0 - \beta_0 + \beta_1\mathbb{E}S - \beta_1\mathbb{E}S + \beta_1\delta \\
&= \beta_1\delta
\end{aligned}
$$

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

---

- We extend this result to the mean counterfactual outcomes under the nonparametric model $\mathcal{M}$.

## A causal inference perspective

- Consider a data structure: $(Y_s, s \in \mathcal{S})$.
- To posit a linear model, let $Y_s = \beta_0 + \beta_1 s + \epsilon_s$ for $s \in \mathcal{S}$, 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 $s' \in \mathcal{S}$, may be expressed

$$\begin{aligned}\mathbb{E}Y_{s'+\delta} - \mathbb{E}Y_{s'} &= [\beta_0 + \beta_1(s' + \delta) + \mathbb{E}\epsilon_{s'+\delta}] - [\beta_0 + \beta_1 s' + \mathbb{E}\epsilon_{s'}] \\ &= \beta_1 \delta\end{aligned}$$

- 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 $\beta_1$.

---

- Note that this analysis is exactly what we're told we **cannot** do in "linear models 101" — that is, the slope of a regression line cannot be interpreted as *causing* a change in the outcome.
- We extend this result to the mean counterfactual outcomes under the nonparametric model $\mathcal{M}$.

## Slope in a semiparametric model

- Consider the stochastic intervention $g^{\star}(\cdot \mid L)$:

$$\mathbb{E}Y_{g^{\star}} = \int_L \int_s \mathbb{E}(Y \mid S = s, L) g(s - \delta \mid L) \cdot ds \cdot dP_0(L)$$
$$= \int_L \int_z \mathbb{E}(Y \mid S = z + \delta, L) g(z \mid L) \cdot dz \cdot dP_0(L),$$

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

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

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

## References


Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24(1):49–64.

Breslow, N., McNeney, B., Wellner, J. A., et al. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *The Annals of Statistics*, 31(4):1110–1139.

Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009). Improved horvitz–thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statistics in Biosciences*, 1(1):32–49.

Díaz, I. and Hejazi, N. S. (2020). Causal mediation analysis for stochastic interventions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):661–683.

Díaz, I., Hejazi, N. S., Rudolph, K. E., and van der Laan, M. J. (2020). Non-parametric efficient causal mediation with intermediate confounders. *Biometrika*.

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.

Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154.

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.

Haneuse, S. and Rotnitzky, A. (2013). Estimation of the effect of interventions that modify the received treatment. *Statistics in medicine*, 32(30):5260–5277.

Hejazi, N. S., Rudolph, K. E., van der Laan, M. J., and Díaz, I. (2020). Nonparametric causal mediation analysis for stochastic interventional (in)direct effects. under review at *Journal of the American Statistical Association (Theory & Methods)*.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

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.

Kennedy, E. H. (2018). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, (just-accepted).

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J. (2013). Direct and indirect effects. *arXiv preprint arXiv:1301.2300*.

Plotkin, S. A. and Gilbert, P. B. (2012). Nomenclature for immune correlates of protection after vaccination. *Clinical Infectious Diseases*, 54(11):1615–1617.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440.

Price, B. L., Gilbert, P. B., and van der Laan, M. J. (2018). Estimation of the optimal surrogate based on a randomized trial. *Biometrics*, 74(4):1271–1281.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155.

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.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

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.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

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).

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Young, J. G., Hernán, M. A., and Robins, J. M. (2014). Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic methods*, 3(1):1–19.