Evaluating treatment efficacy in vaccine clinical trials with two-phase designs using stochastic-interventional effects

Nima Hejazi

Thursday, December 01, 2022

Department of Biostatistics,
T.H. Chan School of Public Health,
Harvard University

nshejazi
nhejazi
nimahejazi.org

Applied Biostatistics Seminar,
Massachusetts General Hospital

Joint work with P.B. Gilbert (Fred Hutch & UW)

Immune Correlates of HIV-1 and COVID-19

The Fights Against HIV-1 and COVID-19

- The HIV-1 epidemic:
  - 1.5 million new infections occurring annually worldwide;
  - new infections outpace patients starting antiretroviral therapy;
  - HIV Vaccine Trials Network's (HVTN) *505* trial evaluated a novel antibody boost vaccine (Hammer et al. 2013).
- The COVID-19 ~~epi~~ ~~pan~~ endemic (Antia and Halloran 2021):
  - ~~270~~ ~~331~~ ~~619~~ 643 million total cases detected globally;
  - new variants emerging, with vaccine uptake globally slowing;
  - COVID-19 Prevention Network's (CoVPN) *COVE* trial focused on Moderna's (mRNA-1273) vaccine (Baden et al. 2021).

Evaluating Vaccines for HIV-1 and COVID-19

- *505*: How would HIV-1 infection risk have differed had the boost vaccine modulated immunogenic responses differently?
- *COVE*: How would COVID-19 disease rate have differed for alternative vaccine-induced immunogenic response profiles?
- **Question**: How can [HIV-1, COVID-19] vaccines be improved through modulating immunogenic response profiles?

Why Measure and Analyze Immune Correlates?

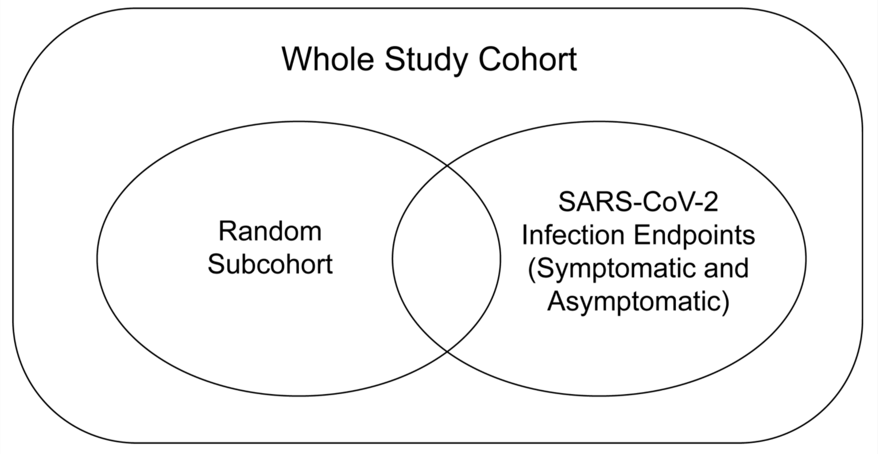
- Two, interrelated goals of vaccine correlates analyses are to
  - identify/validate possible *surrogate endpoints* (Prentice 1989);
  - understand *protective mechanisms* of vaccines.
- If an immune correlate is established to reliably predict VE, subsequent efficacy trials may use it as a primary endpoint.
- This may accelerate the approval of
  - existing vaccines in different populations (e.g., in children);
  - new vaccines within the same class.

Measuring Correlates: Two-Phase Designs

- Often, use case-cohort design (Prentice 1986), a special case of two-phase sampling (Breslow et al. 2003).
- Phase 1: measure baseline, vaccination, endpoint on everyone.
- Phase 2: given baseline, vaccine, endpoint, select members of immune response subcohort with (possibly known) probability.
    - *505*: second-phase sample with 100% of HIV-1 cases and matching of non-cases ($n = 189$ per Janes et al. 2017)).
    - *COVE*: stratified random subcohort ($n \approx 1600$) and all SARS-CoV-2 infection and COVID-19 disease endpoints.

A Simple Two-Phase Design: Case-Cohort

Assaying >30k samples is expensive, statistically unnecessary.



Case-cohort design, per Prentice (1986), as applied to COVE.

Two-phase Sampling Masks the Complete Data Structure

- Complete (unobserved) data $X = (L, A, S, Y) \sim P_0^X \in \mathcal{M}$:
  - L (baseline covariates): sex, age, BMI, behavioral HIV risk,
  - A (treatment): randomized assignment to vaccine/placebo,
  - S (exposure): immune response profile for relevant markers,
  - Y (outcome of interest): infection status at trial's end.
- Observed data $O = (B, BX) = (L, B, BS, Y) \sim P_0 \in \mathcal{M}$.
  - $B \in \{0, 1\}$ indicates inclusion in the second-phase sample.
  - $\pi_0 := \mathbb{P}(B = 1 \mid Y, L)$ must be *known by design* or estimated.
  - Implicitly conditioning on the vaccine arm: $O = \{X \mid A = 1\}$.

Causal Effects for Quantitative Exposures

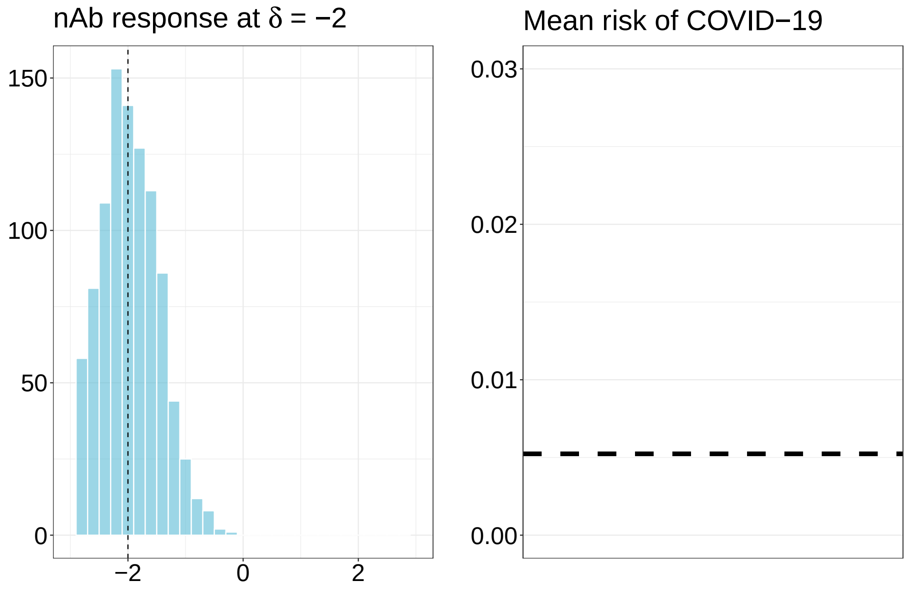
Static Interventions Aren't Enough

- Describe the manner in which X is hypothetically generated by a nonparametric structural equation model (Pearl 2009):

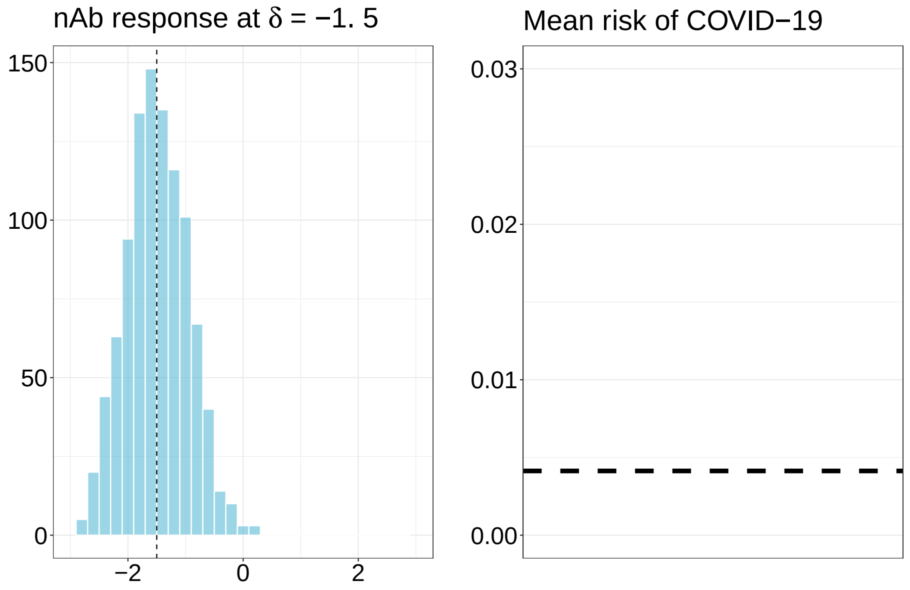
  $$L = f_L(U_L); A \sim \text{Bern}(0.5); S = f_S(A, L, U_S); Y = f_Y(S, A, L, U_Y)$$

- Implies a model for the distribution of counterfactual random variables induced by interventions on the system.
- A *static* intervention replaces f_S with a specific value s in its conditional support $S \mid L$.
- This requires specifying *a priori* a particular value of exposure under which to evaluate the outcome — but what value?

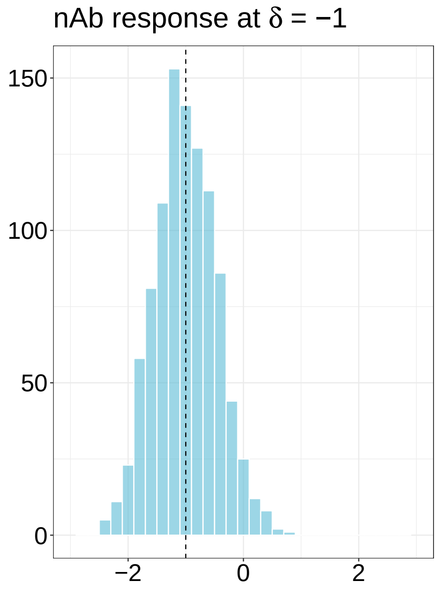
Disease Risk Under Shifted Immunogenic Responses

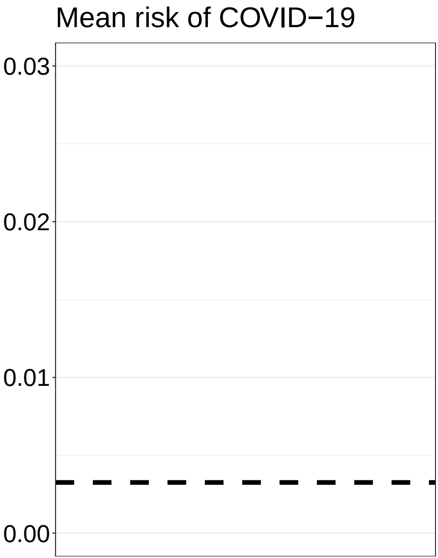


Disease Risk Under Shifted Immunogenic Responses

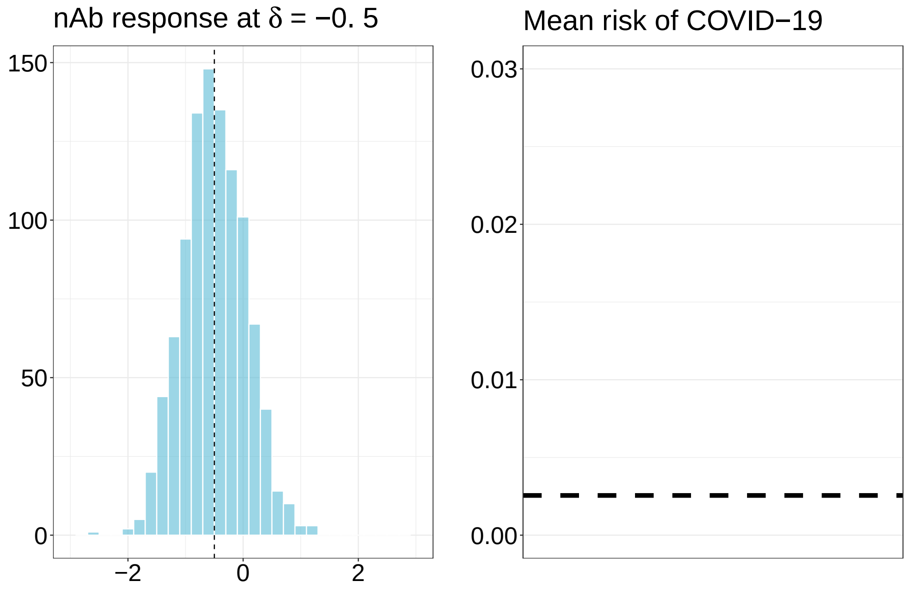


Disease Risk Under Shifted Immunogenic Responses





Disease Risk Under Shifted Immunogenic Responses



Stochastic Interventions Define the Causal Effects of Shifts

- Stochastic interventions modify the value S would naturally assume by *shifting* the natural exposure distribution.
- Díaz and van der Laan (2012; 2018)'s shift interventions[1]

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

- Our estimand is $\psi_{0,\delta} := \mathbb{E}_{P_0^\delta}\{Y_{d(S,L)}\}$, which is identified by

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

[1] Haneuse and Rotnitzky (2013) introduced modified treatment policies.

Interpreting the Causal Effects of Shift Interventions

- Consider a data structure: $(Y_s, s \in \mathcal{S})$.
- Let $Y_s = \beta_0 + \beta_1 s + \epsilon_s$, 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 $Y_{s'+\delta} - Y_{s'}$ may be expressed (for some $s' \in \mathcal{S}$)

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

- A *unit shift* for $s' \in S$ (i.e., $\delta = 1$) causes a counterfactual difference in Y of magnitude β_1.

Stochastic–Interventional Vaccine Efficacy

- Causal parameter based on vaccine efficacy (VE) estimands:

$$\begin{aligned}\mathrm{SVE}(\delta) &= 1 - \frac{\mathbb{E}[\mathbb{P}(Y = 1 \mid S = d(s, l), A = 1, L = l)]}{\mathbb{P}(Y(0) = 1)} \\ &= 1 - \frac{\psi_{0,\delta}}{\mathbb{P}(Y(0) = 1)}\end{aligned}$$

- $\mathbb{P}(Y(0) = 1)$: counterfactual infection risk in the placebo arm — under randomization, $\mathbb{P}(Y(0) = 1) \equiv \mathbb{P}(Y = 1 \mid A = 0)$.
- Summarizes VE via stochastic interventions across δ, per the CoVPN immune correlates SAP[2] (Gilbert et al. 2021a;b).

[2] SAP published at `https://doi.org/10.6084/m9.figshare.13198595`.

Efficient Estimation in Two-Phase Designs

Estimation of the Counterfactual Mean $\psi_{0,\delta}$

An *estimator* $\psi_{n,\delta}$ of $\psi_{0,\delta} \coloneqq \Psi(P_0)$ is *efficient* if and only if

$$\psi_{n,\delta} - \psi_{0,\delta} = n^{-1} \sum_{i=1}^{n} D^{\star}(P_0)(O_i) + o_P(n^{-1/2}) \ ,$$

where $D^{\star}(P)$ is the *efficient influence function* (EIF) of $\psi_{0,\delta}$ with respect to the nonparametric model $\mathcal{M}$ at a distribution P.

The EIF of $\psi_{0,\delta}$ is indexed by two key *nuisance parameters*

$$\overline{Q}_Y(S, L) \coloneqq \mathbb{E}_P(Y \mid S, L) \qquad \textit{outcome mechanism}$$

$$g_S(S \mid L) \coloneqq p(S \mid L) \qquad \textit{generalized propensity score}$$

Flexible, Efficient, Doubly Robust Estimation

- The efficient influence function of $\psi_{0,\delta}$ with respect to $\mathcal{M}$ is

$$D^{\star}_{F}(P_0)(o) = \frac{g_{0,S}(d^{-1}(s,l) \mid l)}{g_{0,S}(s \mid l)}(y - \overline{Q}_{0,Y}(s,l)) + \overline{Q}_{0,Y}(d(s,l),l) - \psi_{0,\delta}.$$

- The one-step bias-corrected estimator:

$$\psi^{+}_{n} = \frac{1}{n}\sum_{i=1}^{n} \overline{Q}_{n,Y}(d(S_i,L_i),L_i) + D^{\star}_{n}(O_i).$$

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

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

- Both doubly robust: flexible modeling for nuisance estimation.

Augmented Estimators for Two-Phase Sampling Designs

- Rose and van der Laan (2011) suggested inverse probability of censoring weighted (IPCW) loss functions:

$$\mathcal{L}(P_0^X)(O) = \frac{B}{\pi_0(Y, L)} \mathcal{L}(P_0^X)(X)$$

- When the sampling mechanism $\pi_0(Y, L)$ is known by design, this procedure yields a reasonably reliable estimator.
- When data-adaptive regression must be used — that is, when $\pi_0(Y, L)$ is not known by design[3]— this is insufficient.

[3] Sampling of non-cases in HVTN 505 used matching (Janes et al. 2017).

Efficiency and Multiple Robustness (Hejazi et al. 2020)

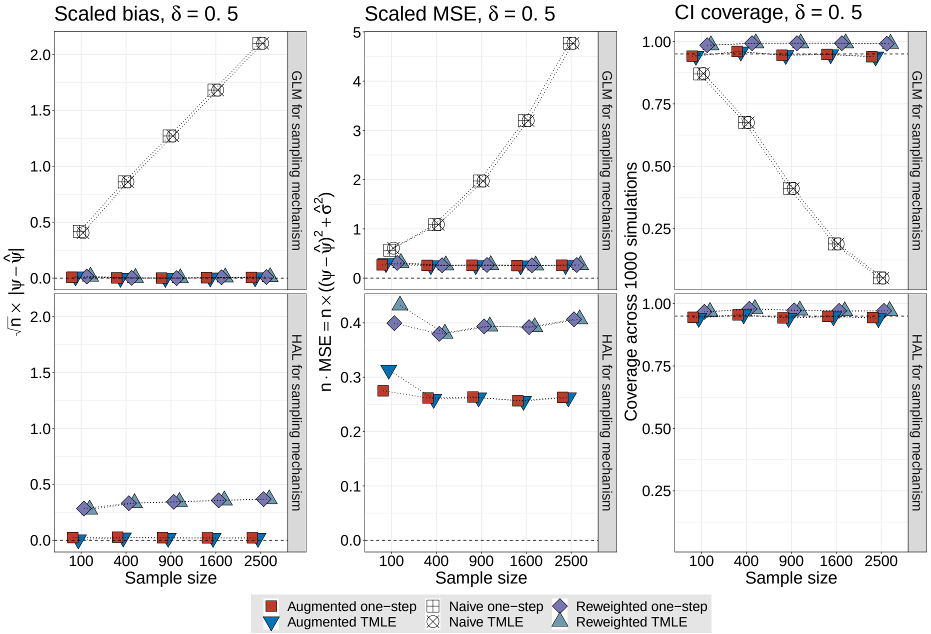
- Then, the IPCW augmentation must be applied to the EIF[4]:

$$D^{\star}(P_0^X)(o) = \frac{b}{\pi_0(y, l)} D_F^{\star}(P_0^X)(x) - \left(1 - \frac{b}{\pi_0(y, l)}\right) \mathbb{E}(D_F^{\star}(P_0^X)(x) \mid B = 1, Y = y, L = l).$$

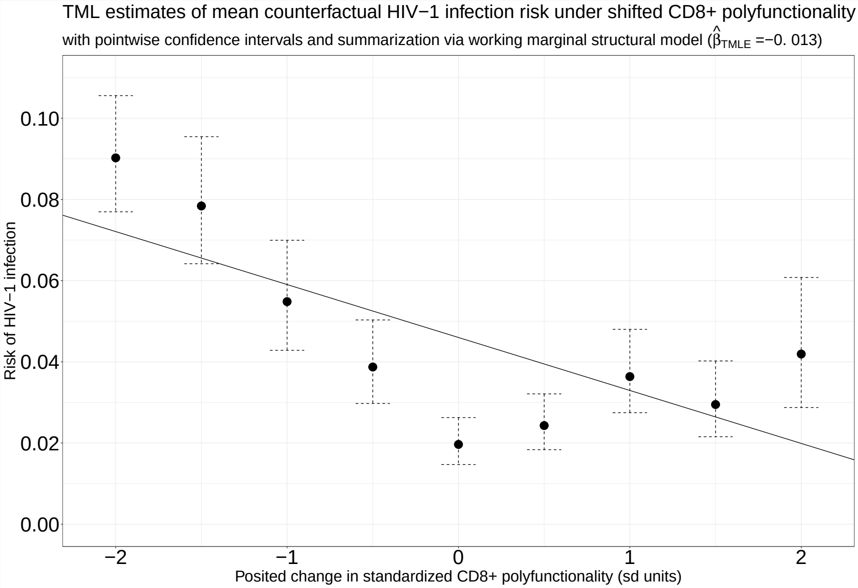
- Expresses observed data EIF $D^{\star}(P_0^X)(o)$ via complete data EIF $D_F^{\star}(P_0^X)(x)$; inclusion of second term improves efficiency.
- An emergent multiple robustness property — combinations of $\{g_0(S \mid L), \overline{Q}_0(S, L)\} \times \{\pi_0(Y, L), \mathbb{E}(D_F^{\star}(P_0^X)(x) \mid B = 1, Y, L)\}$.
- Our `txshift` `R` package implements our estimators of $\psi_{0,\delta}$.

[4] A very general version appears to have been presented in Robins et al. (1994).

Comparing Reweighted and Augmented Estimators

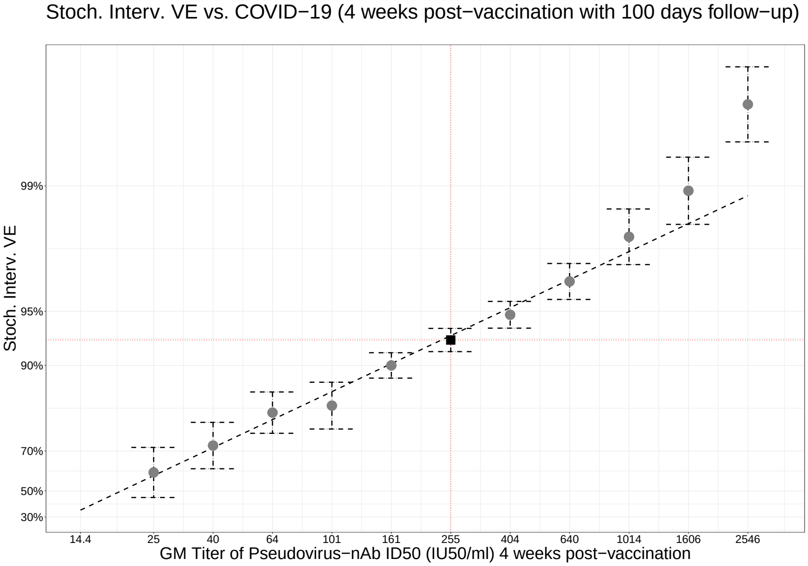


SVE Prediction of HIV-1 Risk thru CD8+ Immune Response



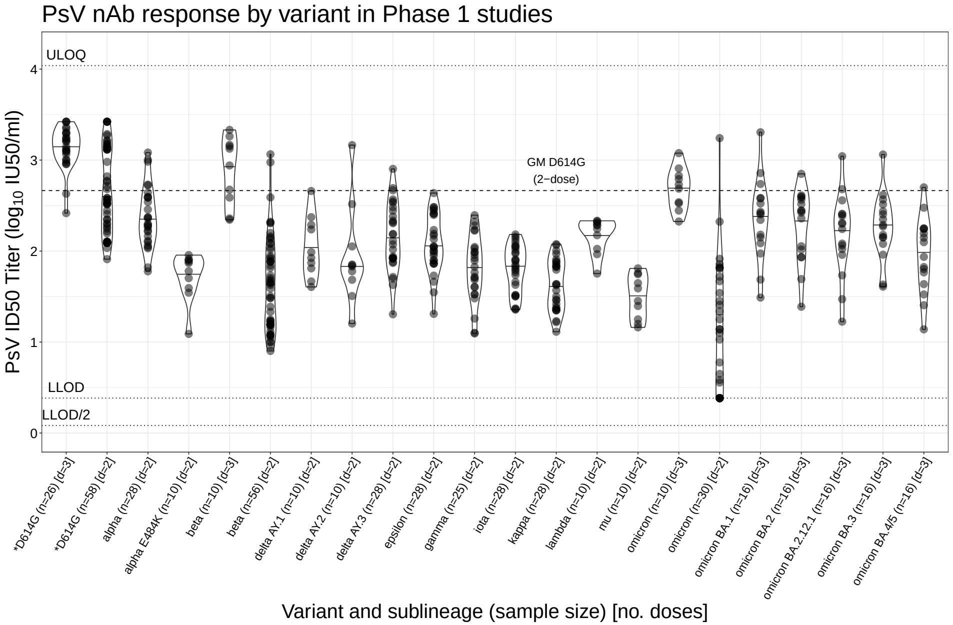
HIV-1 risk change across CD8+ response (`txshift` R package).

SVE Prediction of mRNA-1273 VE thru PsV nAb Correlate

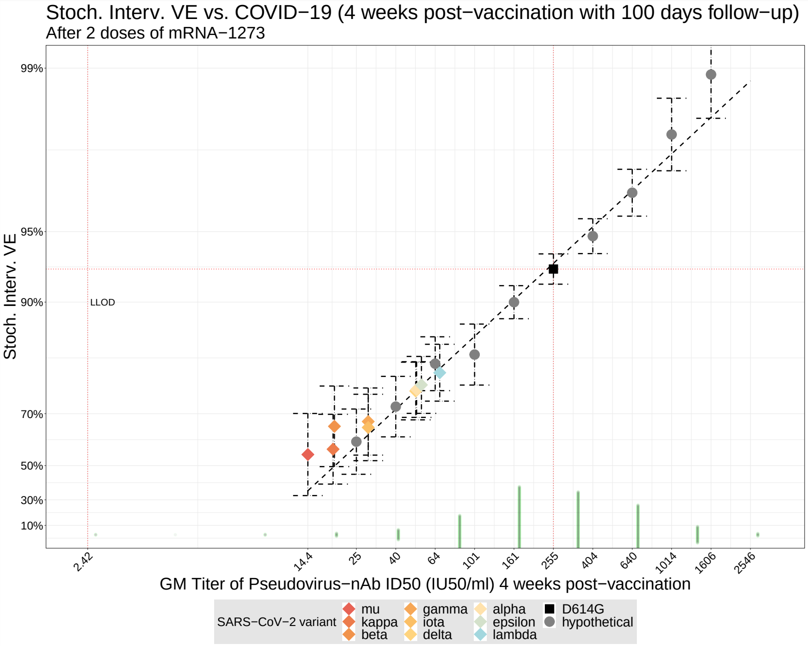


Bridging VE Using Immune Correlates

Pooled Phase 1 Studies: PsV nAb Responses Across Variants



SVE Bridging of mRNA-1273 VE thru PsV nAb Correlate

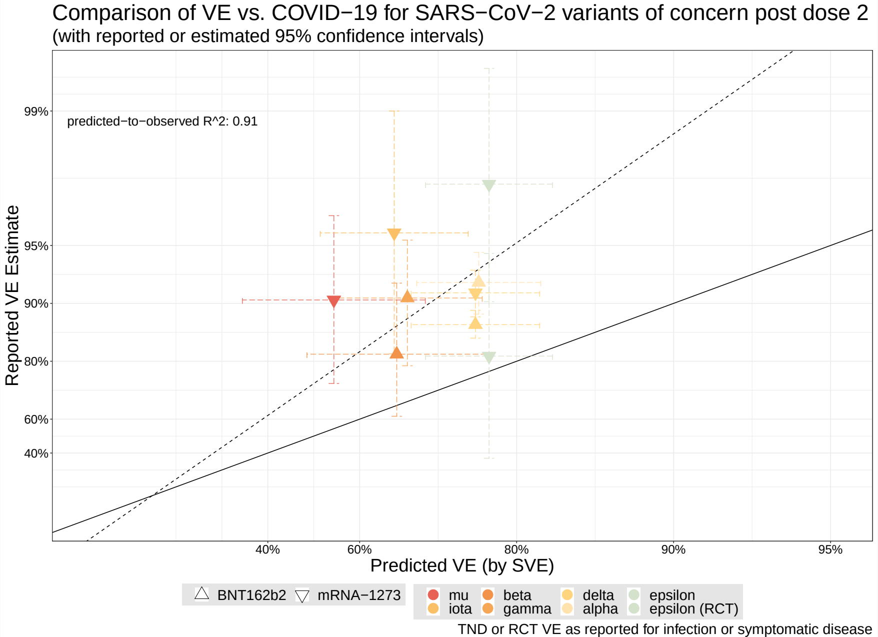


SVE Predictions vs. Real-World Reported Estimates

- Compared δ-calibrated SVE predictions to reported VE estimates, from TND studies or RCTs.
- Inclusion/exclusion criteria for TND-based VE estimates[5]:
  - VE estimated by direct measurement of SARS-CoV-2 variants.
  - Reported VE estimates for mRNA vaccines (BNT162b2 or mRNA-1273), studying VE 2–6 months post-2$^{\text{nd}}$ dose.
  - Allowed flexibility in choice of dosing interval, with some studies extending to 12 weeks between doses.
- Studied concordance of SVE predictions and reported (TND or RCT) estimates of VE following most recent vaccine dose.

[5] Comparison of TND studies performed in collaboration with Dr. Lindsay Carpp.

Concordance of SVE Predictions and Reported VE Estimates



Summary of SVE Prediction for Immunobridging

- SVE prediction shows sharp changes in VE with shifts to the GM titer of the PsV nAb correlate in vaccinees.
- Bridging VE across variants indicates VE drops but stabilizes at 50%, if the model based on ancestral D614G strain holds.
- Post-2$^{\text{nd}}$ dose: For most variants (excepting omicron), the VE estimate ranges from 50% (mu) to 80% (epsilon).
- SVE predictions and real-world VE estimates well-correlated, but SVE predictions may be underestimates, as PsV nAb correlate is an *imperfect* causal *mediator* of total VE.

The Big Picture

- Stochastic interventions provide a framework for formulating novel policies based on natural treatment conditions.
- These modified treatment policies address causal questions about *realistic* interventions on quantitative treatments.
- Large-scale vaccine trials rely upon two-phase designs — but need to (very carefully!) adjust for the resultant sampling bias.
- Efficient estimators with double/multiple robustness can safely answer such questions *while* incorporating machine learning.
- Open source software for such statistical analyses is critical for the methods to have any impact on real-world studies.

Thank you

Thanks for listening. Any questions?

https://nimahejazi.org

https://github.com/nhejazi

https://twitter.com/nshejazi

Appendix

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.

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 (Cox 1958)*
- $Y^{d(s,l)} = Y$ *in the event* $S = d(s, l)$*, for* $i = 1, \ldots, n$

Assumption 2: *No Unmeasured Confounding*

$S \perp\!\!\!\perp Y^{d(s,l)} \mid L = l$, *for* $i = 1, \ldots, n$

Assumption 3: *Positivity*

$s \in \mathcal{S} \implies d(s, l) \in \mathcal{S}$ *for all* $l \in \mathcal{L}$*, where* $\mathcal{S}$ *denotes the support of* S *conditional on* $L = l$ *for all* $i = 1, \ldots n$

Literature: Díaz and van der Laan (2012; 2018)

- *Proposal:* Evaluate outcome under an altered *intervention distribution* — e.g., $P_\delta(g_{0,S})(S = s \mid L) = g_{0,S}(s - \delta(L) \mid L)$.
- Identification conditions for a statistical parameter of the counterfactual outcome $\psi_{0,\delta}$ under such an intervention.
- Show that the causal quantity of interest $\mathbb{E}_{P_0^\delta}\{Y_{d(S,L)}\}$ is identified by a functional of the distribution of O, i.e.,

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

Literature: Haneuse and Rotnitzky (2013)

- *Proposal:* Characterization of stochastic interventions as *modified treatment policies* (MTPs).
- Assumption of *piecewise smooth invertibility* allows for the post-intervention distribution of any MTP to be recovered:

$$g_{0,S}(s \mid l; \delta) = \sum_{j=1}^{J(l)} \mathbb{I}_{\delta,j}\{h_j(s,l), l\} g_0\{h_j(s,l) \mid l\} h_j'(s,l)$$

- MTPs account for the natural value of exposure S yet may be interpreted as imposing an altered intervention mechanism.

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.
- Linear model: consider $Y_i = \beta_0 + \beta_1 S_i + \epsilon_i$, with error $\epsilon_i \sim N(0, 1)$.
- Letting δ be a change in S, $Y_{S+\delta} - Y_S$ may be expressed

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

- So, a *unit shift* in S (i.e., $\delta = 1$) induces a change in the difference in outcomes of magnitude β_1.

Slope in a Semiparametric Model

- Consider the stochastic intervention $g_\delta(\cdot \mid L)$:

$$\begin{aligned}\mathbb{E}Y_{g_\delta} &= \int_L \int_s \mathbb{E}(Y \mid S = s, L) g(s - \delta \mid L) ds dP_0(L) \\ &= \int_L \int_z \mathbb{E}(Y \mid S = z + \delta, L) g(z \mid L) dz dP_0(L),\end{aligned}$$

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

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

$$\begin{aligned}\mathbb{E}Y_{g_\delta} - \mathbb{E}Y &= \int_L \int_z [\mathbb{E}(Y \mid S = z + \delta, L) - \mathbb{E}(Y \mid S = z, L)] \\ &\qquad g(z \mid L) dz dP_0(L) \\ &= [\beta(z + \delta) + \theta(L)] - [\beta z + \theta(L)] \\ &= \beta\delta\end{aligned}$$

Flexible Conditional Density Estimation of $g_{0,S}$

- Díaz and van der Laan (2011)'s conditional density estimator:

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

  - Re-expressed as hazard regressions in repeated measures data.
  - Tuning parameter $t \approx$ bandwidth in kernel density estimation.

- When càdlàg (RCLL) with finite sectional variation, we have

$$\operatorname{logit}\{\mathbb{P}(s \in [\alpha_{t-1}, \alpha_t) \mid L)\} = \beta_0 + \sum_{w \subset \{1,\ldots,d\}} \sum_{i=1}^{n} \beta_{w,i}\phi_{w,i},$$

for appropriate basis functions $\{\phi_{w,i}\}_{i=1}^{n}$ (Gill et al. 1995).

Flexible Conditional Density Estimation of $g_{0,S}$

- Utilizing a particular basis construction for ϕ_w, van der Laan (2017)'s HAL estimator achieves $n^{-1/4}$ convergence rate[6].
- Loss-based cross-validation allows selection of a suitable HAL estimator, which has only the ℓ_1 regularization term λ:

$$\beta_{n,\lambda} = \underset{\beta : |\beta_0| + \sum_{w \subset \{1,\ldots,d\}} \sum_{i=1}^{n} |\beta_{w,i}| < \lambda}{\arg\min} P_n \mathcal{L}(g_{\beta,\lambda,S}),$$

where $\mathcal{L}(\cdot)$ is an appropriate loss function, giving $\{\lambda_n, \beta_n\}$.

- We denote by $g_{n,\lambda,S} \coloneqq g_{\beta_{n,\lambda},S}$, the HAL estimate of $g_{0,S}$.
- Our `haldensify` `R` package implements our estimator of $g_{0,S}$.

[6] Similar rates can be achieved via *local* (vs. global) smoothness assumptions on $g_{n,S}$ (see, e.g., Robins et al. 2008, Mukherjee et al. 2017, Liu et al. 2021).

A Useful Class of Functions

Consider space of *cadlag* functions with *finite variation norm*.

Def. cadlag = *left-hand continuous* with *right-hand limits*

Variation norm Let $\theta_s(u) = \theta(u_s, 0_{s^c})$ be the *section* of θ that sets the coordinates in *s equal to zero*.

The *variation norm* of θ can be written:

$$|\theta|_v = \sum_{s \subset \{1,\ldots,d\}} \int | \, d\theta_s(u_s) \, |,$$

where $x_s = (x(j) : j \in s)$ and the sum is over all subsets.

Variation Norm

We can represent the function θ as

$$\theta(x) \quad = \quad \theta(0) + \sum_{s \subset \{1,\ldots,d\}} \int \mathbb{I}(x_s \geq u_s) d\theta_s(u_s),$$

For discrete measures $d\theta_s$ with *support points* $\{u_{s,j} : j\}$ we get a *linear combination* of indicator *basis functions*:

$$\theta(x) = \theta(0) + \sum_{s \subset \{1,\ldots,d\}} \sum_j \beta_{s,j} \theta_{u_{s,j}}(x),$$

where $\beta_{s,j} = d\theta_s(u_{s,j})$, $\theta_{u_{s,j}}(x) = \mathbb{I}(x_s \geq u_{s,j})$, and

$$|\theta|_v = \theta(0) + \sum_{s \subset \{1,\ldots,d\}} \sum_j |\beta_{s,j}|.$$

Convergence Rate of HAL

We have, for $\alpha(d) = 1/(d+1)$,

$$|\theta_{n,M} - \theta_{0,M}|_{P_0} = o_P(n^{-(1/4+\alpha(d)/8)}).$$

Thus, if we select $M > |\theta_0|_v$, then

$$|\theta_{n,M} - \theta_0|_{P_0} = o_P(n^{-(1/4+\alpha(d)/8)}) \ .$$

Due to oracle inequality for the cross-validation selector M_n,

$$|\theta_{n,M_n} - \theta_0|_{P_0} = o_P(n^{-(1/4+\alpha(d)/8)}) \ .$$

Improved convergence rate (Bibaut and van der Laan 2019):

$$|\theta_{n,M_n} - \theta_0|_{P_0} = o_P(n^{-1/3} \log(n)^{d/2}) \ .$$

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 ϵ 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 Ψ_n of the target parameter, defining update $\overline{Q}_n^{\star}$ of the initial estimate $\overline{Q}_{n,\epsilon_n}$:
$$\Psi_n = \Psi(P_n^{\star}) = \frac{1}{n}\sum_{i=1}^{n} \overline{Q}_n^{\star}(d(S_i, L_i), L_i).$$

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 Ψ_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 TMLE in each iteration.

References

Antia, R. and Halloran, M. E. (2021). Transition to endemicity: Understanding covid-19. *Immunity*, 54(10):2172–2176.

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., McGettigan, J., Khetan, S., Segall, N., Solis, J., Brosz, A., Fierro, C., Schwartz, H., Neuzil, K., Corey, L., Gilbert, P., Janes, H., Follmann, D., Marovich, M., Mascola, J., Polakowski, L., Ledgerwood, J., Graham, B. S., Bennett, H., Pajon, R., Knightly, C., Leav, B., Deng, W., Zhou, H., Han, S., Ivarsson, M., Miller, J., Zaks, T., and the COVE Study Group (2021). Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine*, 384(5):403–416.

Bibaut, A. F. and van der Laan, M. J. (2019). Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm. *arXiv preprint arXiv:1907.09244*.

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.

Cox, D. R. (1958). *Planning of Experiments*. Wiley.

Díaz, I. and van der Laan, M. J. (2011). Super learner based conditional density estimation with application to marginal structural models. *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.

Gilbert, P. B., Fong, Y., Benkeser, D., Andriesen, J., Borate, B., Carone, M., Carpp, L. N., Díaz, I., Fay, M. P., Fiore-Gartland, A., Hejazi, N. S., Huang, Y., Huang, Y., Hyrien, O., Janes, H. E., Juraska, M., Li, K., Luedtke, A., Nason, M., Randhawa, A. K., van der Laan, L., Williamson, B. D., Zhang, W., and Follmann, D. (2021a). Covpn covid-19 vaccine efficacy trial immune correlates statistical analysis plan.

Gilbert, P. B., Montefiori, D. C., McDermott, A. B., Fong, Y., Benkeser, D., Deng, W., Zhou, H., Houchens, C. R., Martins, K., Jayashankar, L., Castellino, F., Flach, B., Lin, B. C., O'Connell, S., McDanal, C., Eaton, A., Sarzotti-Kelsoe, M., Lu, Y., Yu, C., Borate, B., van der Laan, L. W., Hejazi, N. S., Huynh, C., Miller, J., El Sahly, H. M., Baden, L. R., Baron, M., De La Cruz, L., Gay, C., Kalams, S., Kelley, C. F., Kutner, M., Andrasik, M. P., Kublin, J. G., Corey, L., Neuzil, K. M., Carpp, L. N., Pajon, R., Follmann, D., Donis, R. O., Koup, R. A., and on behalf of the Immune Assays; Moderna, Inc.; Coronavirus Vaccine Prevention Network (CoVPN) / Coronavirus Efficacy (COVE); and United States Government (USG) / CoVPN Biostatistics Teams (2021b). Immune correlates analysis of the mRNA-1273 COVID-19 vaccine efficacy clinical trial. *Science*.

Gill, R. D., van der Laan, M. J., and Wellner, J. A. (1995). Inefficient estimators of the bivariate survival function for three models. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 545–597.

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., van der Laan, M. J., Janes, H. E., Gilbert, P. B., and Benkeser, D. C. (2020). Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials. *Biometrics*.

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.

Liu, L., Mukherjee, R., Robins, J. M., and Tchetgen Tchetgen, E. (2021). Adaptive estimation of nonparametric functionals. *Journal of Machine Learning Research*, 22(99):1–66.

Mukherjee, R., Newey, W. K., and Robins, J. M. (2017). Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*.

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

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.

Robins, J. M., Li, L., Tchetgen Tchetgen, E., and van der Vaart, A. W. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

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.

van der Laan, M. J. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *International Journal of Biostatistics*, 13(2).