

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


Nima Hejazi

Monday, December 19, 2022


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



HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

 nshejazi

 nhejazi

 nimahejazi.org

CMStats: “Effect estimation in various contexts”,
King’s College London

Joint work with P.B. Gilbert & D. Benkeser

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 (?).
- The COVID-19 epidemic endemic (?):
 - 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 (?).

Evaluating Vaccines for HIV-1 and COVID-19

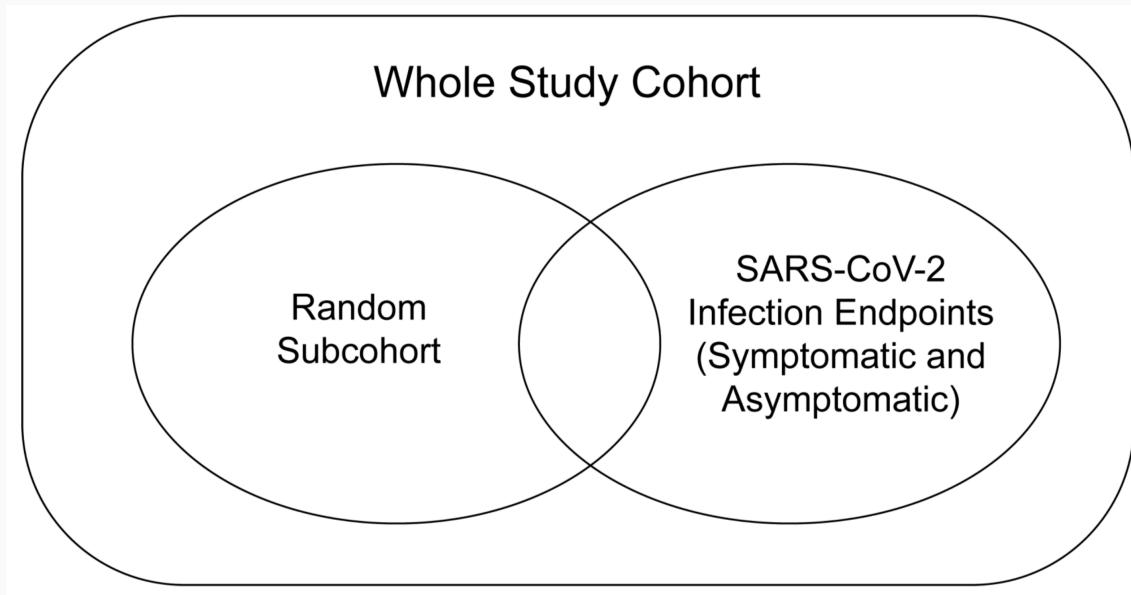
- *505*: How would HIV-1 infection risk have differed had the boost vaccine modulated immunogenic responses differently?
- *COVE*: How would COVID-19 disease rate have differed for alternative vaccine-induced immunogenic response profiles?
- **Question**: How can [HIV-1, COVID-19] vaccines be improved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

Measuring Correlates: Two-Phase Designs

- Often, use case-cohort design (?), a special case of two-phase sampling (?).
- 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 ?).
 - *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 ?, 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: $O = \{X \mid A = 1\}$.

Static Interventions Aren't Enough

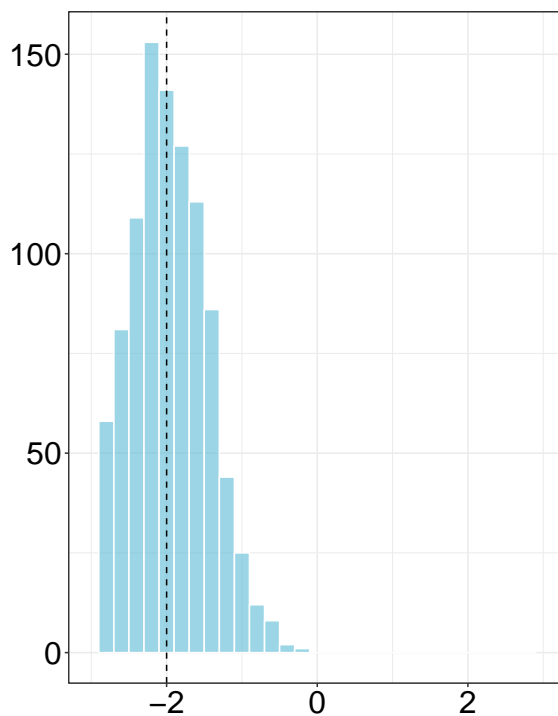
- Describe the manner in which X is hypothetically generated by a nonparametric structural equation model (?):

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

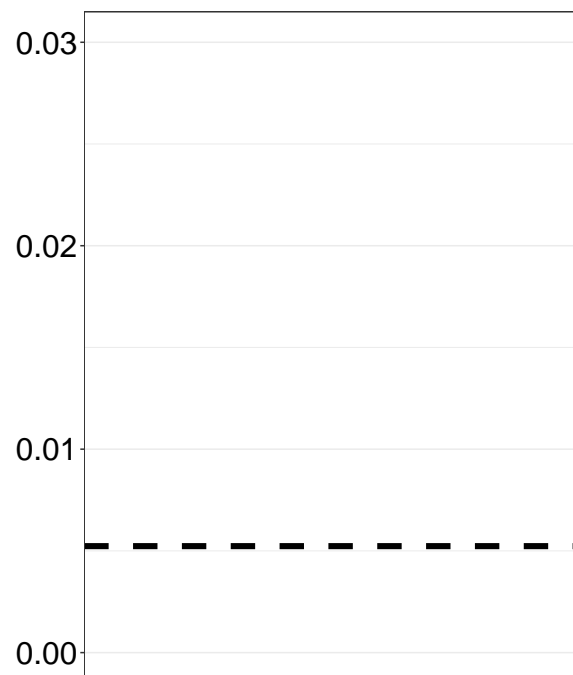
- Implies a model for the distribution of counterfactual random variables induced by interventions on the system.
- A *static* intervention replaces f_S with a specific value s in its conditional support $S | L$.
- This requires specifying *a priori* a particular value of exposure under which to evaluate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

Disease Risk Under Shifted Immunogenic Responses

nAb response at $\delta = -2$

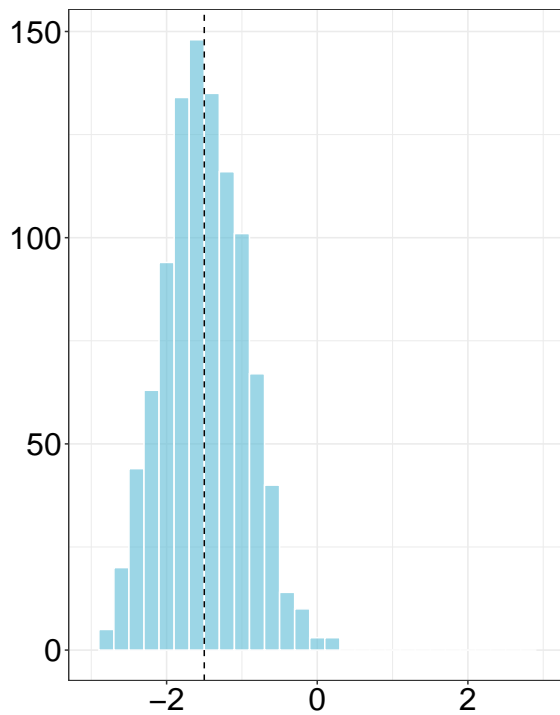


Mean risk of COVID-19

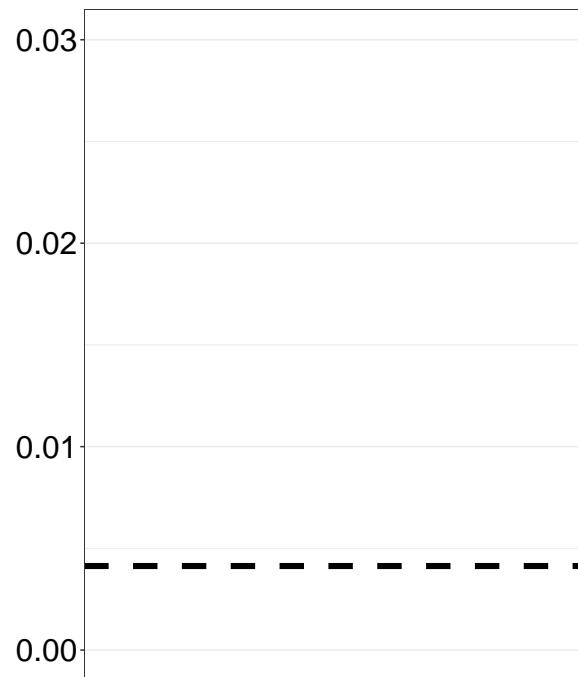


Disease Risk Under Shifted Immunogenic Responses

nAb response at $\delta = -1.5$

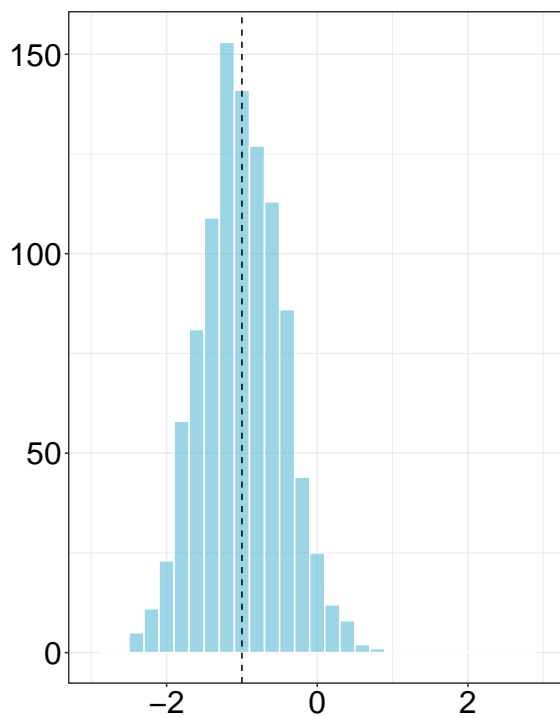


Mean risk of COVID-19

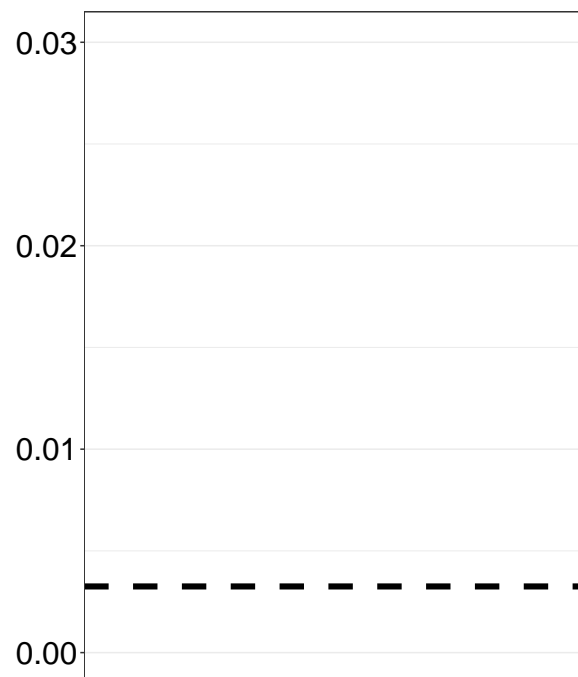


Disease Risk Under Shifted Immunogenic Responses

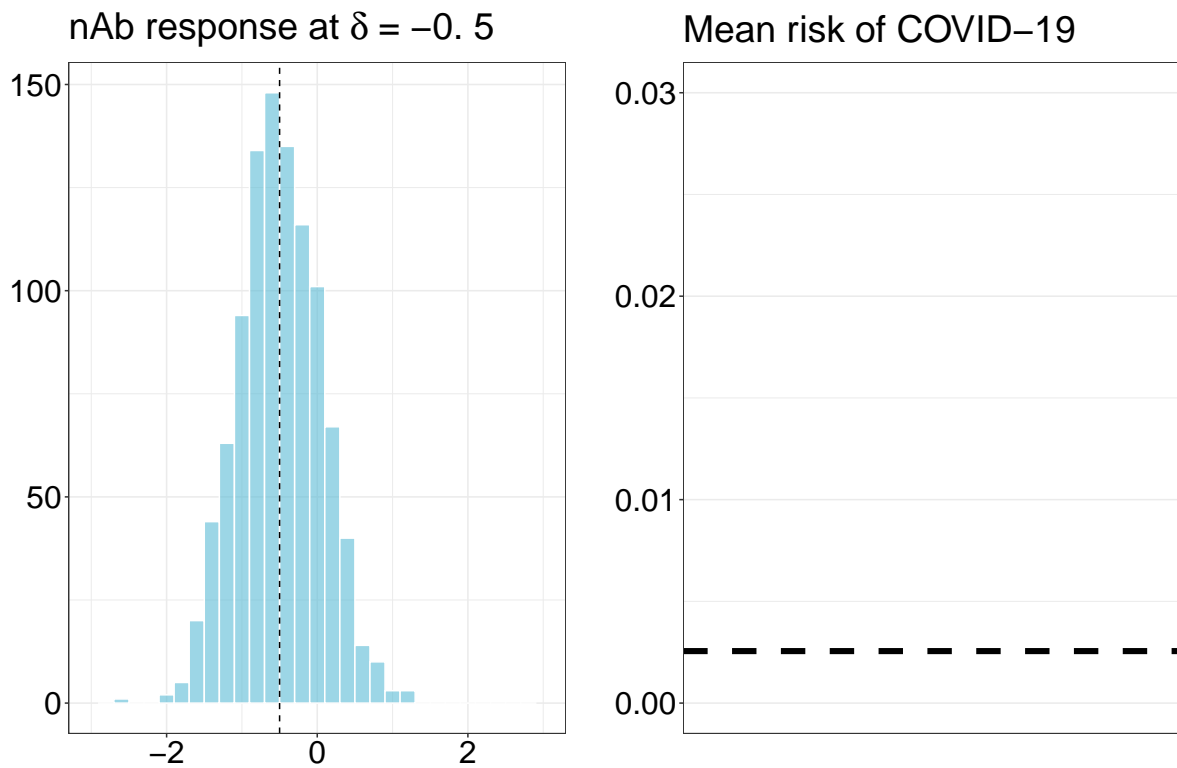
nAb response at $\delta = -1$



Mean risk of COVID-19



Disease Risk Under Shifted Immunogenic Responses



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 $\epsilon_s \sim N(0, \sigma_s^2) \forall s \in \mathcal{S}$.
- 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 $s' \in \mathcal{S}$ (i.e., $\delta = 1$) causes a counterfactual difference in Y of magnitude β_1 .

Stochastic Interventions Define the Causal Effects of Shifts

- Stochastic interventions modify the value S would naturally assume by *shifting* the natural exposure distribution.
- ??'s shift interventions¹

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

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

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

¹? introduced modified treatment policies.

Stochastic–Interventional Vaccine Efficacy

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

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

- $\mathbb{P}(Y(0) = 1)$: 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 SAP² (??).

²SAP published at <https://doi.org/10.6084/m9.figshare.13198595>.

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

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

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

where $D^*(P)$ 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*

$$\begin{aligned} \bar{Q}_Y(S, L) &:= \mathbb{E}_P(Y | S, L) && \text{outcome mechanism} \\ g_S(S | L) &:= p(S | L) && \text{generalized propensity score} \end{aligned}$$

Flexible, Efficient, Doubly Robust Estimation

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

$$D_F^*(P_0)(o) = \frac{g_{0,S}(d^{-1}(s, l) | l)}{g_{0,S}(s | l)} (y - \bar{Q}_{0,Y}(s, l)) + \bar{Q}_{0,Y}(d(s, l), l) - \psi_{0,\delta}.$$

- The one-step bias-corrected estimator:

$$\psi_n^+ = \frac{1}{n} \sum_{i=1}^n \bar{Q}_{n,Y}(d(S_i, L_i), L_i) + D_n^*(O_i).$$

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

$$\psi_n^* = \frac{1}{n} \sum_{i=1}^n \bar{Q}_{n,Y}^*(d(S_i, L_i), L_i).$$

- Both doubly robust: flexible modeling for nuisance estimation.

Augmented Estimators for Two-Phase Sampling Designs

- ? 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 design³— this is insufficient.

³Sampling of non-cases in HVTN 505 used matching (?).

Efficiency and Multiple Robustness (?)

- Then, the IPCW augmentation must be applied to the EIF⁴:

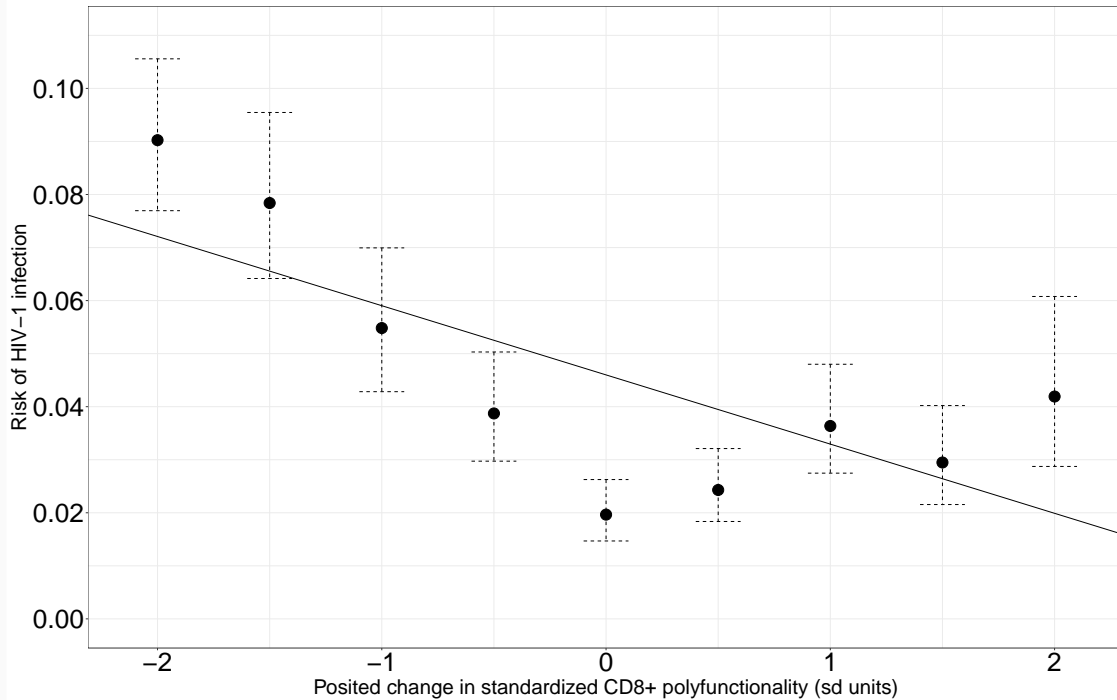
$$D^*(P_0^X)(o) = \frac{b}{\pi_0(y, l)} D_F^*(P_0^X)(x) - \left(1 - \frac{b}{\pi_0(y, l)}\right) \mathbb{E}(D_F^*(P_0^X)(x) \mid B = 1, Y = y, L = l).$$

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

⁴A very general version appears to have been presented in ?.

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

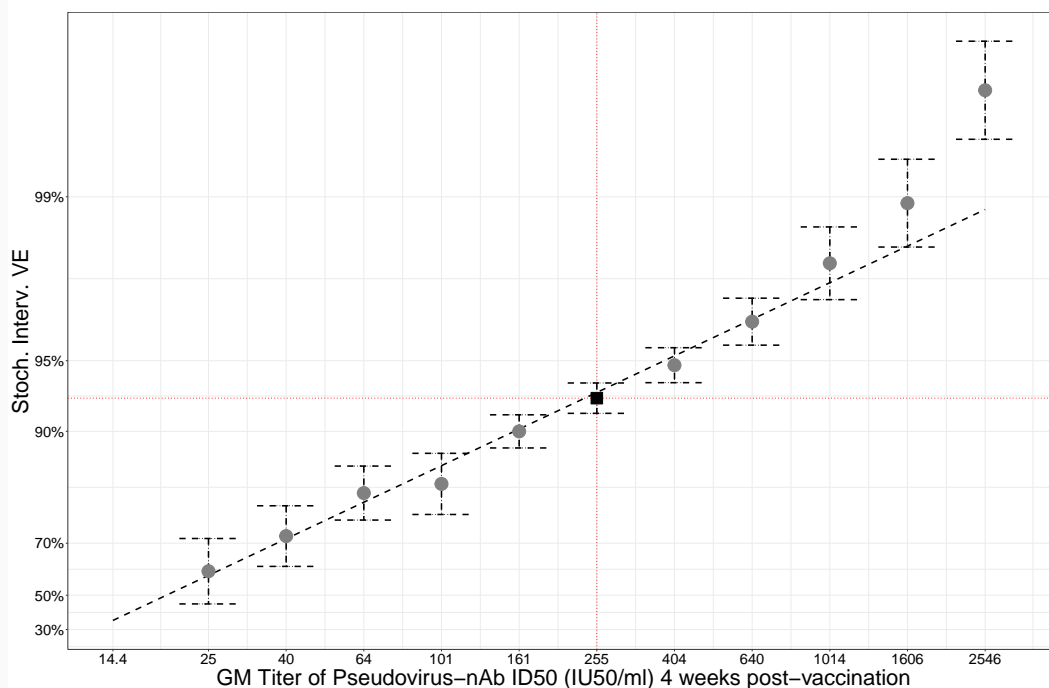
TML estimates of mean counterfactual HIV-1 infection risk under shifted CD8+ polyfunctionality with pointwise confidence intervals and summarization via working marginal structural model ($\hat{\beta}_{\text{TMLE}} = -0.013$)



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

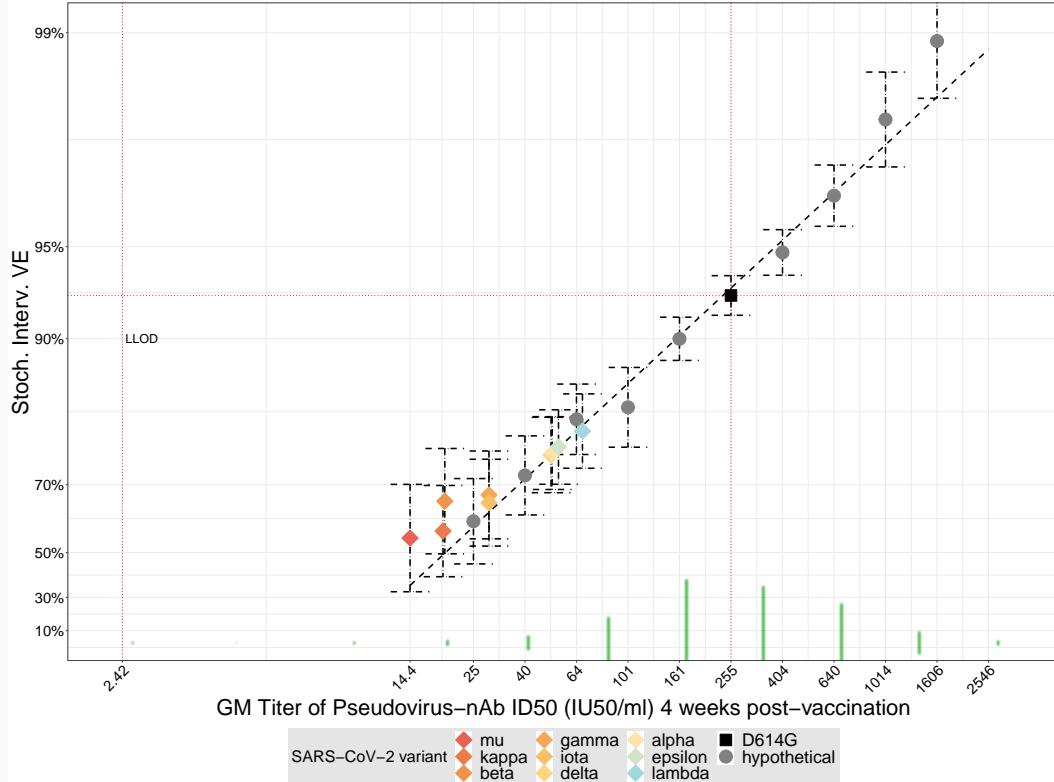
SVE Prediction of mRNA-1273 VE thru PsV nAb Correlate

Stoch. Interv. VE vs. COVID-19 (4 weeks post-vaccination with 100 days follow-up)



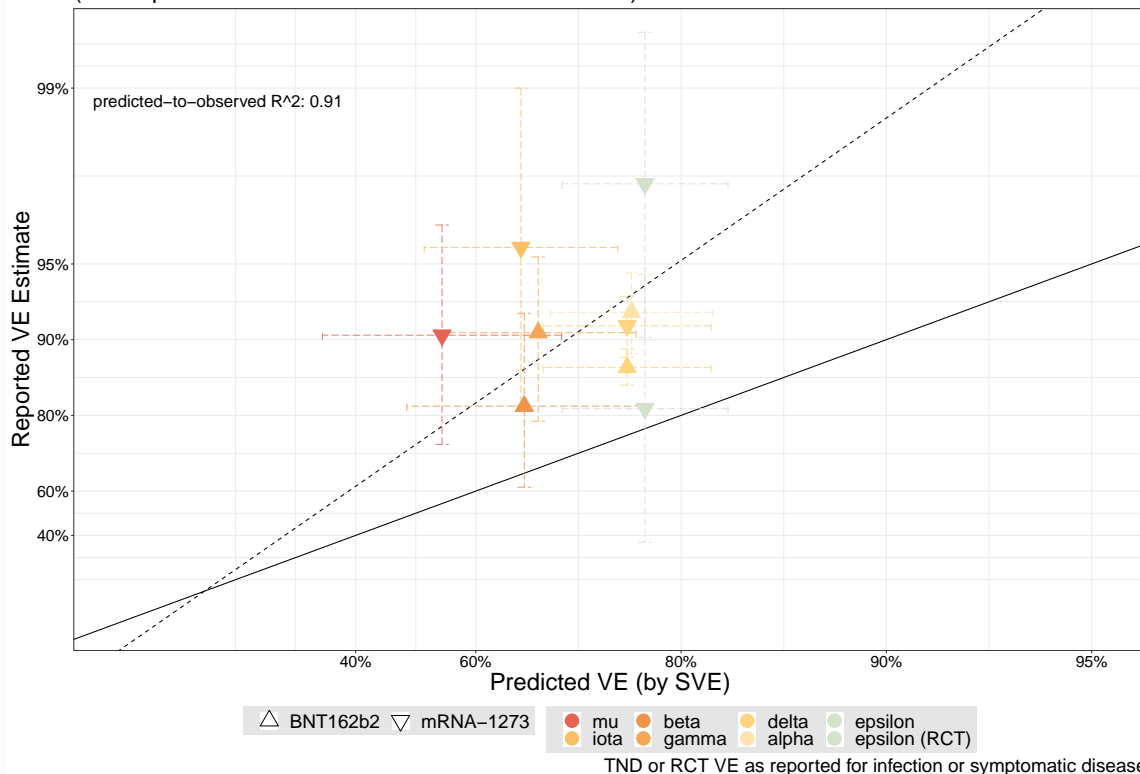
SVE Bridging of mRNA-1273 VE thru PsV nAb Correlate

Stoch. Interv. VE vs. COVID-19 (4 weeks post-vaccination with 100 days follow-up)
After 2 doses of mRNA-1273



Concordance of SVE Predictions and Reported VE Estimates

Comparison of VE vs. COVID-19 for SARS-CoV-2 variants of concern post dose 2
(with reported or estimated 95% confidence intervals)



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 (?)

- 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 (?) — 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 $i = 1, \dots, n$ and $j \neq i$, or lack of interference (?)
- $Y^{d(s, l)} = Y$ in the event $S = d(s, l)$, for $i = 1, \dots, n$

Assumption 2: *No Unmeasured Confounding*

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

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 $i = 1, \dots, n$

Literature: ??

- *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 $\mathbb{E}_{P_0} \{ Y_{d(S, L)} \}$ 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: ?

- *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:

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

- 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 $g_\delta(\cdot | L)$:

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

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

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

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

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

- 's conditional density estimator:

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

- 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

$$\text{logit}\{\mathbb{P}(s \in [\alpha_{t-1}, \alpha_t] | L)\} = \beta_0 + \sum_{w \in \{1, \dots, d\}} \sum_{i=1}^n \beta_{w,i} \phi_{w,i},$$

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

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

- Utilizing a particular basis construction for ϕ_w , ?'s HAL estimator achieves $n^{-1/4}$ convergence rate⁵.
- Loss-based cross-validation allows selection of a suitable HAL estimator, which has only the ℓ_1 regularization term λ :

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

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

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

⁶Similar rates can be achieved via *local* (vs. global) smoothness assumptions on $g_{n,S}$ (see, e.g., ???).

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:

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

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

Variation Norm

We can represent the function θ as

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

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

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

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

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

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 (?):

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

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).