Robust Nonparametric Inference for Stochastic Interventions Under Multi-Stage Sampling

Nima S. Hejazi, Mark J. van der Laan, and David C. Benkeser

Group in Biostatistics & Department of Statistics, University of California, Berkeley
Department of Biostatistics and Bioinformatics, Emory University

Overview & Motivations

1. We consider the problem of efficiently estimating the effect of a stochastic shift interventions for problem settings in which multi-stage sampling complicates the observed data structure.
2. We present a novel approach: an augmented targeted maximum likelihood estimator of a parameter defined as the outcome under a stochastic intervention with
   - consistency and efficiency guarantees even under multi-stage sampling, and
   - a form of multiple double robustness inherited from its constituent parts.
3. The proposed nonparametric estimation procedure provably attains fast convergence rates even when incorporating machine learning estimators.
4. A recent software implementation — the "*txshift*" R package [2] — has been developed for applying this methodology in complete generality, including for causal inference and variable importance analyses.

Data: HIV Vaccine Trials

- We illustrate the utility of our approach by applying the new method and software in an investigation of the effects of immune response biomarkers on HIV vaccine efficacy.
- *Question of interest:* **How does risk of HIV infection differ under posited shifts of the distribution of an immune response in the vaccine arm of an efficacy trial?**
- We simulate a data structure based on the HVTN 505 HIV-1 efficacy trial, as in [3]:
  - About 2500 participants, with all observed cases matched to controls.
  - Background (W): sex, age, BMI, etc.
  - Intervention (A): immunobiomarkers (i.e., T-Cell profiles from ICS assays on preserved HIV-1-stimulated PBMCs).
  - Outcome (Y): HIV-1 infection status.
- *Takeaway:* **Variable importance measure for ranking multiple immune responses by their utility as immunogenicity study endpoints in future HIV-1 vaccine trials.**

Methodology I: The Effect of a Stochastic Intervention

- Consider $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, with no assumptions placed on the statistical model $\mathcal{M}$.
- Rather than a deterministic intervention, consider a shift of the treatment (i.e., instead of $A = a$, consider a shift of the intervention so that $A = a + \delta$ for an aribtrary δ).
- As a comparison with the general linear model, the shift δ may be thought of as a part of the nonparametric analog to the slope of a regression line — i.e., $\beta^{\text{NP}}_{\text{slope}} = \frac{\mathbb{E}[Y \mid A+\delta] - \mathbb{E}[Y \mid A]}{\delta^2}$.
- To protect against positivity violations, make the shifting mechanism a function of the observed data: $d(a, w) = a + \delta$, if $a + \delta < u(w)$ and $d(a, w) = a$ otherwise.

We consider a simple causal target parameter, introduced in [4]:

$$\Psi(P) = \mathbb{E}_{\text{P}} \overline{Q}(d(A, W), W), \tag{1}$$

for which the efficient influence function (EIF), given in [1], is

$$D(P)(o) = H(a, w)y - \overline{Q}(a, w) + \overline{Q}(d(a, w), w) - \Psi(P), \tag{2}$$

where the auxiliary term, $H(a, w)$, takes the form $H(a, w) = \mathbb{I}(a < u(w)) \frac{g_0(a - \delta \mid w)}{g_0(a \mid w)} + \mathbb{I}(a \geq u(w) - \delta)$.

We obtain Wald-style inference via the limiting distribution: $\sqrt{n}(\Psi_n - \Psi) \rightarrow N(0, \text{Var}(D(P_0)))$.

Methodology II: Corrections for Multi-Stage Sampling

- In the HVTN 505 HIV-1 trial, all infected individuals are matched to controls using a complex mechanism, which makes the observed data structure $O = (W, \Delta A, Y)$, where
  - $V = (W, Y)$ is the set of variables defining the sampling mechanism,
  - $\Delta = f(V) \in \{0, 1\}$ is the missingness mechanism introduced by sampling,
  - $\Pi_0(V) = \mathbb{P}(\Delta = 1 \mid V)$, letting $\Pi_n(V)$ be an estimator of $\Pi_0(V)$.
- An IPCW-TMLE, introduced by [5], augments the loss function with IPC weights to overcome the problem introduced by sampling: $\mathcal{L}(P_X)(O) = \frac{\Delta}{\Pi_n(V)} \mathcal{L}^F(P_X)(X)$, for full data $X = (W, A, Y)$.
- When working in a nonparametric model, the efficient influence function estimating equation is complexified by sampling: $0 = P_n \frac{\Delta}{\Pi^*_n(V)} D^F(P^*_{X,n}) - \left\{ \frac{\Delta}{\Pi^*_n(V)} - 1 \right\} \mathbb{E}_n(D^F(P^0_{X,n}) \mid \Delta = 1, V)$.
- Fortuitously, this augmented estimator exhibits a unique form of *multiple double robustness* — through combinations of the terms (g, Q) and $(\Pi, \mathbb{E}_0(D^F(P^F) \mid V))$.

Results & Discussion

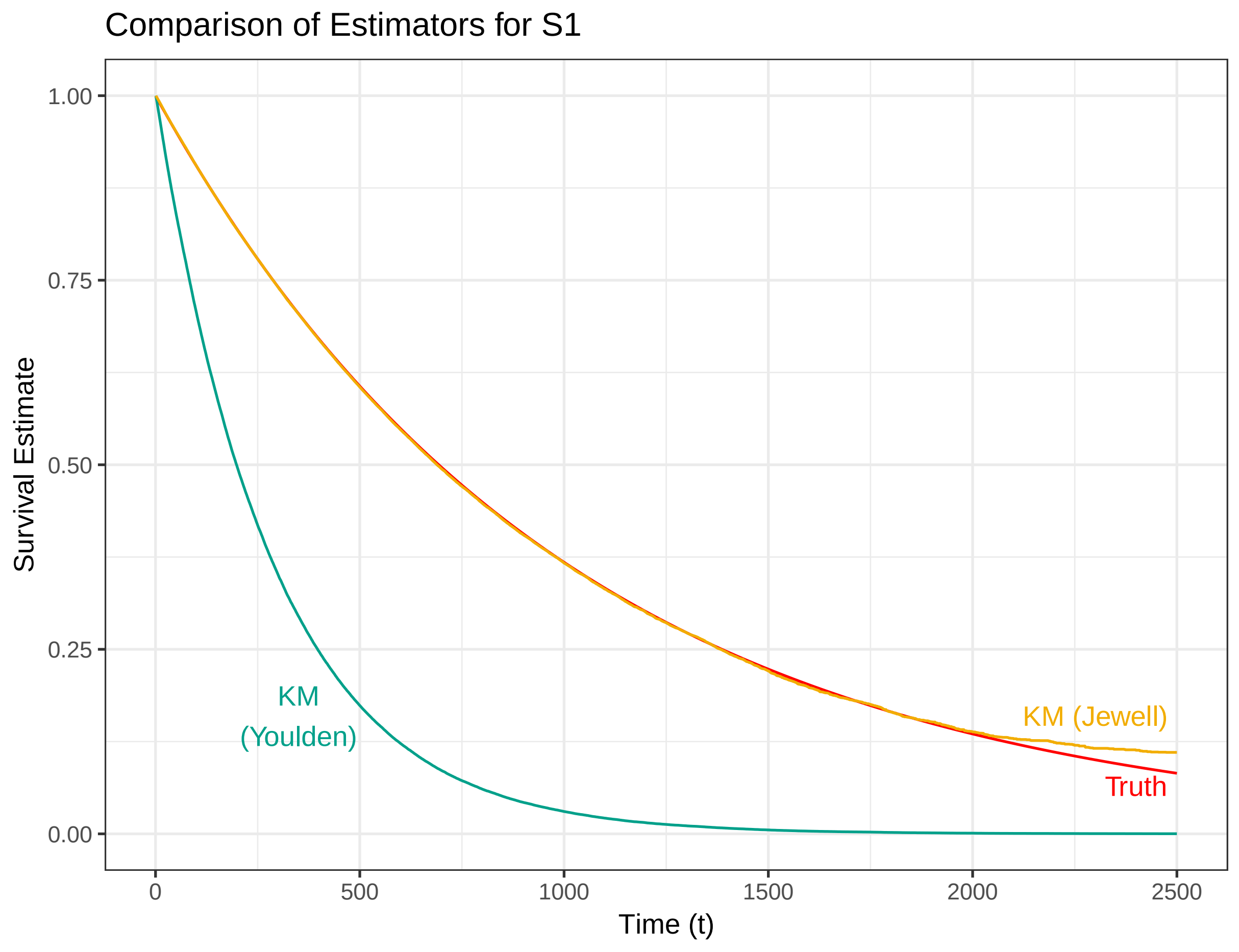

Figure 1: This figure demonstrates...

- ...
- ...
- ...

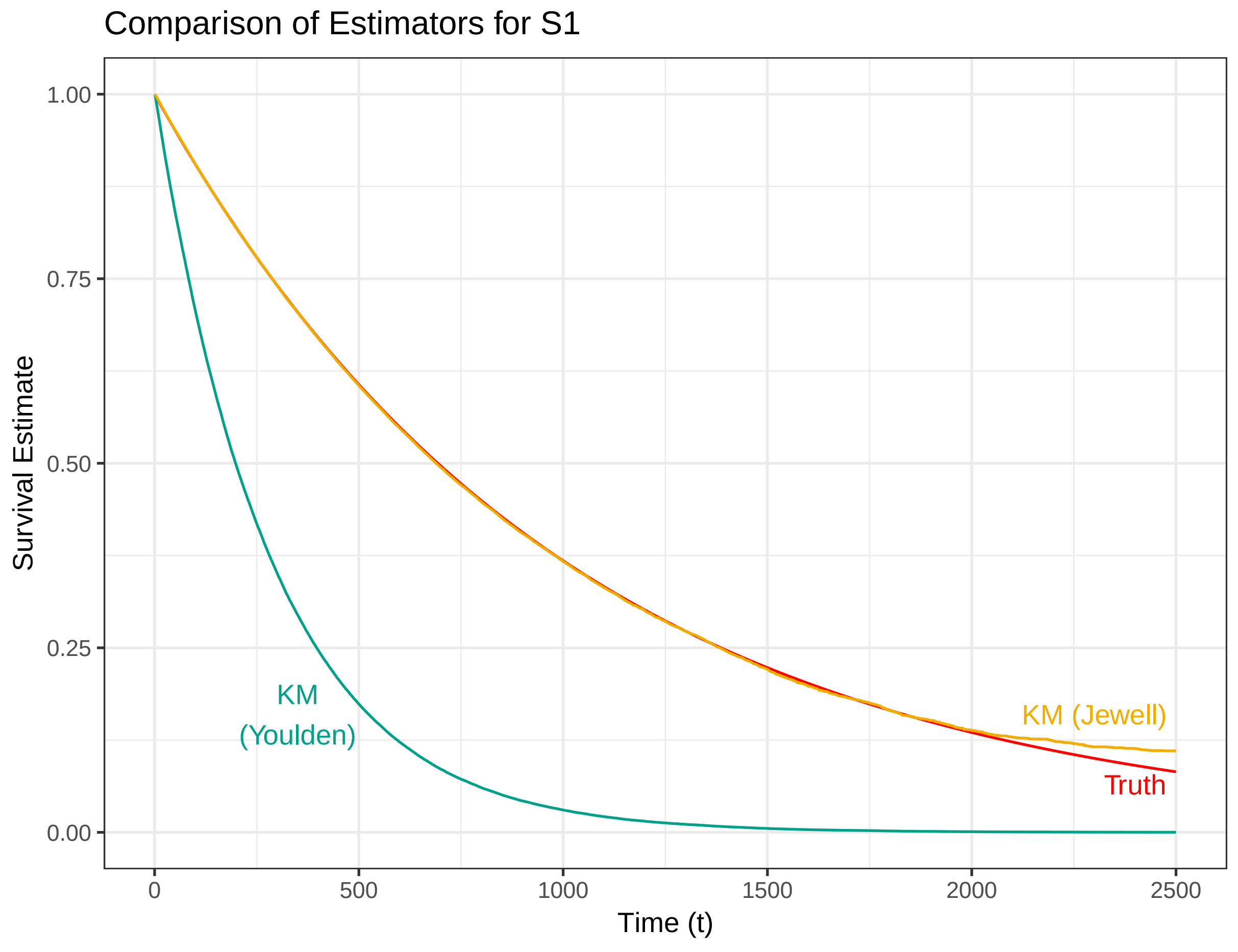

Figure 2: This figure demonstrates...

- ...
- ...
- ...

Principal References

[1] Iván Díaz and Mark J van der Laan. Stochastic treatment regimes. In *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pages 167–180. Springer Science & Business Media, 2018.

[2] Nima S Hejazi, Mark J van der Laan, and David C Benkeser. *txshift: Targeted Minimum Loss-Based Estimation of the Causal Effect of Stochastic Interventions and Variable Importance Analysis.* URL `https://github.com/nhejazi/txshift`. R package version x.x.x.

[3] Holly E Janes, Kristen W Cohen, Nicole Frahm, Stephen C De Rosa, Brittany Sanchez, John Hural, Craig A Magaret, Shelly Karuna, Carter Bentley, Raphael Gottardo, et al. 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, 2017.

[4] Iván Díaz Muñoz and Mark J van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

[5] Sherri Rose and Mark J van der Laan. A targeted maximum likelihood estimator for two-stage designs. *The International Journal of Biostatistics*, 7(1):1–21, 2011.

Contact Information

- **N.S. Hejazi**, Ph.D. student, Group in Biostatistics, NHEJAZI@BERKELEY.EDU
- **M.J. van der Laan**, Professor of Biostatistics & Statistics, LAAN@BERKELEY.EDU
- **D.C. Benkeser**: Assistant Professor of Biostatistics, BENKESER@EMORY.EDU