Targeted Learning of Causal Impacts of Multiple Time-Point Interventions on Survival

Mark van der Laan
Division of Biostatistics, UC Berkeley

November 20, 2019
Department of Statistics
Joint work with Wilson Cai

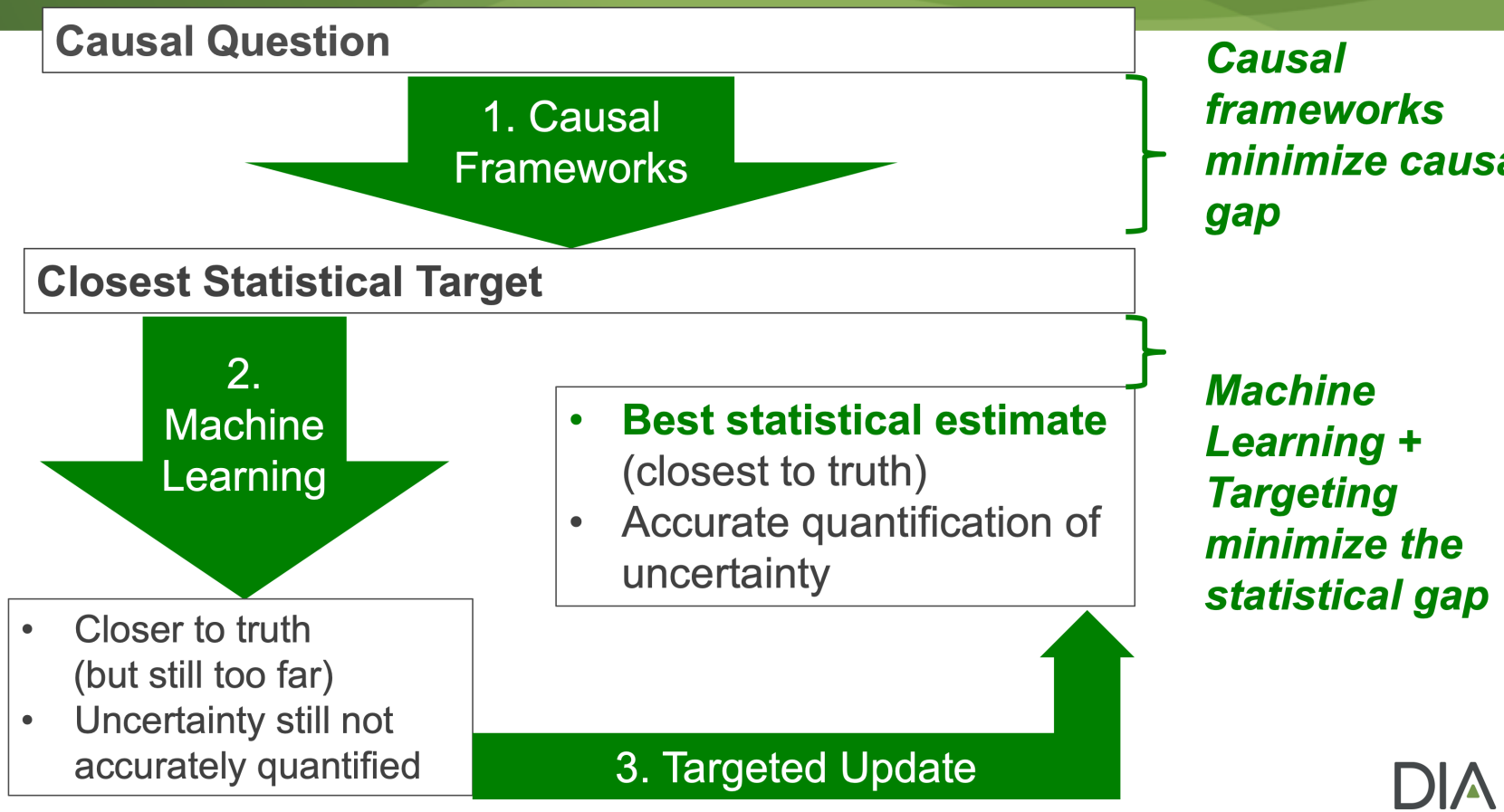
Outline

Roadmap of Statistical Learning

- **Observed data**: Realization of a random variable $O^n = (O_1, \ldots, O_n)$ with a probability distribution (say) P_0^n on n units.
- **Model stochastic system of observed data realistically**: Statistical model $\mathcal{M}^n$ is set of possible probability distributions of the data.
- **Define query about stochastic system**: Function Ψ from model $\mathcal{M}^n$ to real line, where $\Psi(P_0^n)$ is the true answer to query about our stochastic system. **Estimand is chosen so that it best approximates the answer to causal question of interest**.
- **Estimator**: An a priori-specified algorithm that takes the observed data O^n and returns an estimate ψ_n to the *true answer to query*. Benchmarked by a dissimilarity-measure (e.g., MSE) w.r.t true answer to query.
- **Confidence interval for true answer to query:** Establish approximate sampling probability distribution of the estimator (e.g., based on CLT), and corresponding statistical inference.

Outline

Targeted Learning



DIA

Outline

Highly Adaptive Lasso (HAL)

- This is a machine learning algorithm that estimates functionals (e.g outcome regression and propensity score) by approximating them with linear model in many ($\leq n2^d$) tensor product indicator basis functions, constraining the L_1-norm of the coefficient vector, and choosing it with cross-validation (vdL, 2015, Benkeser, vdL, 2017)
- Guaranteed to converge to truth at rate $n^{-1/3}(\log n)^d$ in sample size n (Bibaut, vdL, 2019): only assumption that true function is right-continuous, left-hand limits, and has finite sectional variation norm.
- When used in super-learner library (or by itself), TMLE (targeted learning) is guaranteed **consistent, (double robust) asymptotically normal and efficient**: one only needs to assume *strong positivity assumption*.

Example: HAL-MLE of conditional hazard

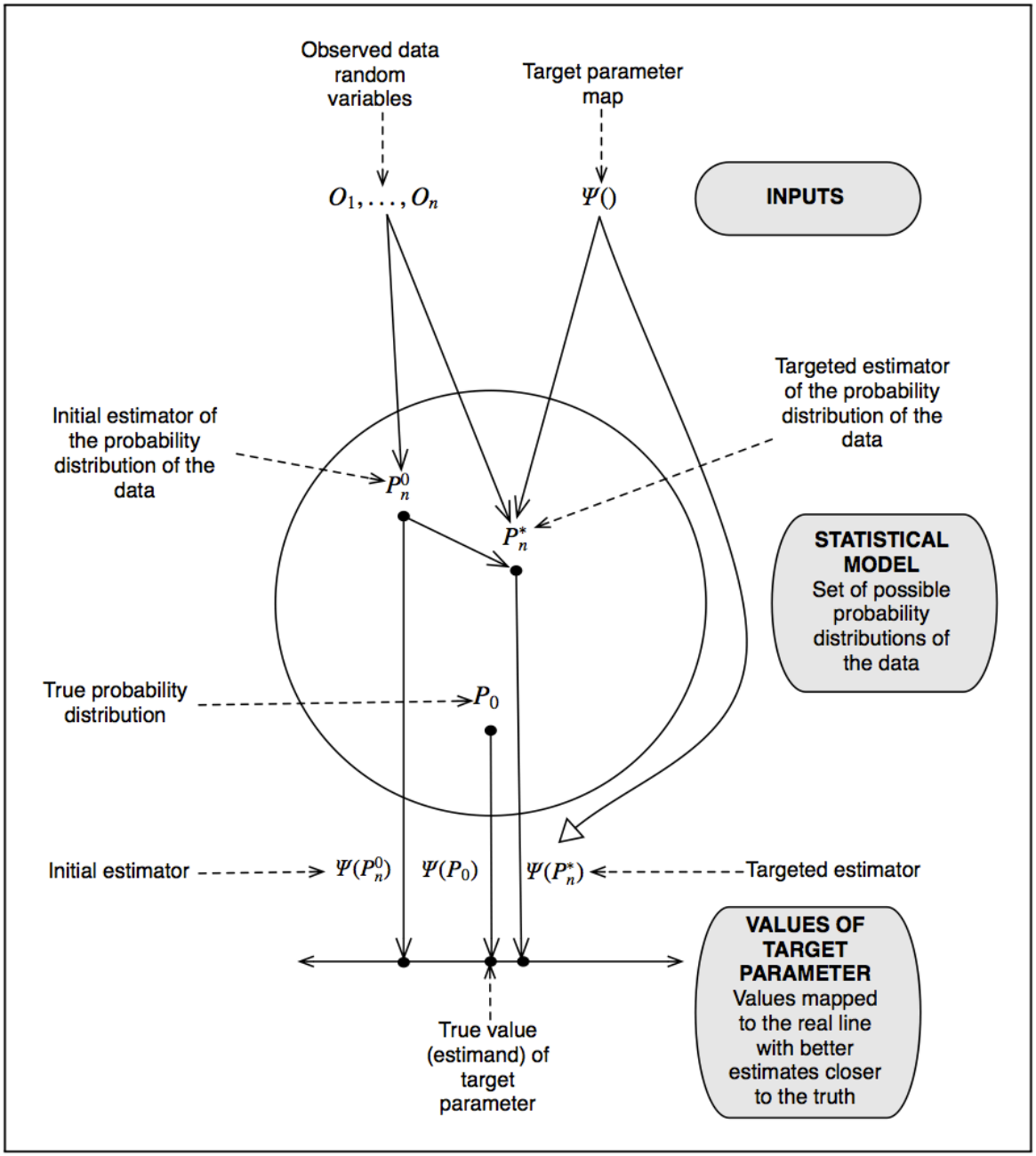
- Suppose that $O = (W, A, \tilde{T} = \min(T, C), \Delta = I(T \leq C))$, and that we are interested in estimating the conditional hazard $\lambda(t \mid A, W)$.
- Let $L(\lambda)$ be the log-likelihood loss.
- If T is continuous, we could parametrize $\lambda(t \mid A, W) = \exp(\psi(t, A, W))$, or, if T is discrete, $\text{Logit}\lambda(t \mid A, W) = \psi(t, A, W)$.
- We can represent $\psi = \sum_{s \subset \{1,\ldots,d\}} \beta_{s,j} \phi_{u_{s,j}}$ as linear combination of indicator basis functions, where L^1-norm of β represents the sectional variation norm of ψ.
- Therefore, we can compute the HAL-MLE of λ with either Cox-Lasso or logistic Lasso regression (glmnet()).

Outline

1. Ingredients of Targeted Learning: Causal Framework, Super-learning, Targeting
2. Highly Adaptive Lasso (HAL) estimator: candidate for Super-Learner library
3. Targeted Minimum Loss Based Estimation (TMLE)
4. Universal Least favorable submodels for one-step TMLE targeting multidimensional target estimands
5. Example: One-step TMLE of causal impact of point intervention on survival
6. TMLE of Causal Effects of Multiple Time Point Interventions on Survival
7. Sequential Regression Representation of Treatment Specific Mean
8. TMLE package for multiple time point interventions in longitudinal studies
9. TMLE in complex observational study of diabetes (Neugebauer et al.)

Outline

Targeted minimum loss based estimation (TMLE)



TMLE

Let $D^*(P)$ be canonical gradient/efficient influence curve of target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ at $P \in \mathcal{M}$.

Initial Estimator: P_n^0 initial estimator of P_0. We recommend super-learning.

Targeting of initial estimator: Construct so called least favorable parametric submodel $\{P_{n,\epsilon}^0 : \epsilon\} \subset \mathcal{M}$ through P_n^0 so that $\frac{d}{d\epsilon} L(P_{n,\epsilon}^0)\Big|_{\epsilon=0}$ spans the canonical gradient $D^*(P_n^0)$ at P_n^0, where (e.g.) $L(P)(O) = -\log p(O)$ is log-likelihood loss. Let

$$\epsilon_n = \arg\min_{\epsilon} \sum_i L(P_{n,\epsilon}^0)(O_i)$$

be the MLE, and $P_n^* = P_{n,\epsilon_n}^0$.

TMLE of ψ_0: The TMLE of ψ_0 is plug-in estimator $\Psi(P_n^*)$.

Solves optimal estimating equation: $P_n D^*(P_n^*) \equiv \frac{1}{n} \sum_i D^*(P_n^*)(O_i) \approx 0$.

Local least favorable submodel

Let $O \sim P_0 \in \mathcal{M}$. Let $\Psi : \mathcal{M} \to \mathbb{R}$ be a one-dimensional target parameter, and let $D^*(P)$ be its canonical gradient at P. A 1-d local least favorable submodel $\{p_\epsilon^{lfm} : \epsilon\}$ satisfies

$$\frac{d}{d\epsilon} \log p_\epsilon^{lfm}) \Big|_{\epsilon=0} = D^*(P).$$

Equivalently, the score of an LFM maximizes the Cramer-Rao lower bound over all 1-d parametric submodels $\{P_\epsilon : \epsilon\}$ through P:

$$CR(h \mid P) = \lim_{\epsilon \to 0} \frac{(\Psi(P_{\epsilon,h}) - \Psi(P))^2}{-2P \log dP_{\epsilon,h}/dP}.$$

That is, an LFM has a local behavior that maximizes the square change in target parameter per change in information.

Outline

Universal least favorable submodel

We define a 1-d universal least favorable submodel at P as a submodel $\{P_\epsilon : \epsilon\}$ so that for all ϵ

$$\frac{d}{d\epsilon} \log \frac{dP_\epsilon}{dP} = D^*(P_\epsilon).$$

This acts as a local least favorable submodel at any point ϵ on its path.

TMLE based on ULFM is a one-step TMLE

Let P_n^0 be an initial estimator of P_0. Suppose that, given a $P \in \mathcal{M}$, we can construct a universal least favorable parametric model $\{P_\epsilon^{ulfm} : \epsilon \in (-a, a)\} \subset \mathcal{M}$. Let

$$\epsilon_n^0 = \arg\max_\epsilon P_n \log \frac{dP_{n,\epsilon}^0}{dP_n^0}.$$

Let $P_n^1 = P_{n,\epsilon_n^0}^0$. Since ϵ_n^0 is a local maximum, P_n^1 solves its score equation, given by $P_n D^*(P_n^1) = 0$. That is, the TMLE is given by $\Psi(P_n^1)$.

Universal least favorable submodel

For $\epsilon \geq 0$, we recursively define

$$p_\epsilon = p \exp\left(\int_0^\epsilon D^*(P_x)dx\right),$$

and, for $\epsilon < 0$, we recursively define

$$p_\epsilon = p \exp\left(-\int_\epsilon^0 D^*(P_x)dx\right).$$

Universal LFM in terms of local LFM

One can also define it in terms of a given local LFM p_ϵ^{lfm}: for $\epsilon > 0$ and $d\epsilon > 0$, we have

$$p_{\epsilon+d\epsilon} = p^{\mathrm{lfm}}_{\epsilon,d\epsilon}.$$

That is, $p_{\epsilon+d\epsilon}$ equals the local LFM $\{p_\delta^{lfm} : \delta\}$ through $p = p_\epsilon$ at local value $\delta = d\epsilon$. Similarly, we define it for $\epsilon < 0$.

A universal canonical submodel that targets a multidimensional target parameter

Let $\Psi(P) = (\Psi(P)(t) : t)$ be multidimensional (e.g., infinite dimensional).
Let $D^*(P) = (D_t^*(P) : t)$ be the vector-valued efficient influence curve.
Consider the following recursively defined submodel: for $\epsilon \geq 0$, we define

$$\begin{aligned} p_\epsilon &= p\Pi_{[0,\epsilon]} \left(1 + \frac{\{P_n D^*(P_x)\}^\top D^*(P_x)}{\| D^*(P_x) \|} dx \right) \\ &= p \exp \left(\int_0^\epsilon \frac{\{P_n D^*(P_x)\}^\top D^*(P_x)}{\| D^*(P_x) \|} dx \right). \end{aligned}$$

Theorem:

We have $\{p_\epsilon : \epsilon \geq 0\}$ is a family of probability densities, its score at ϵ is a linear combination of $D_t^*(P_\epsilon)$ for $t \in \tau$, and is thus in the tangent space $T(P_\epsilon)$, and we have

$$\frac{d}{d\epsilon} P_n \log(p_\epsilon) = \| P_n D^*(P_\epsilon) \| .$$

As a consequence, we have $\frac{d}{d\epsilon} P_n L(P_\epsilon) = 0$ implies $\| P_n D^*(P_\epsilon) \| = 0$. Under regularity conditions, we also have $\{p_\epsilon : \epsilon\} \subset \mathcal{M}$.

One-step TMLE of multi-dimensional target parameter

Let $p_n^0 \in \mathcal{M}$ be an initial estimator of p_0. Let $\epsilon_n = \arg\max_\epsilon P_n \log p_\epsilon$. Let $p_n^* = p_{n,\epsilon_n}^0$ and $\psi_n^* = \Psi(P_n^*)$. We have

$$P_n D^*(P_n^*) = 0.$$

Outline

One-step TMLE of treatment specific survival curve

We investigated the performance of one-step TMLE for treatment specific survival curve based on $O = (W, A, \tilde{T} = \min(T, C), \Delta = I(T \leq C))$.

Data structure

- dynamic treatment intervention: $W \rightarrow d(W)$.
- $S_d(t)$ is defined by

$$\Psi(P)(t) = E_P\left[P\left(T > t | A = d(W), W\right)\right]$$

- Focus on $d(W) = 1$.

Efficient influence curve

The efficient influence curve for $\Psi(P)(t)$ is (Hubbard et al., 2000)

$$D_t^*(P) = \sum_{k \leqslant t} h_t(g_A, S_{A_c}, S)(k, A, W) \Big[I(\widetilde{T} = k, \Delta = 1) -$$

$$I(\widetilde{T} \geqslant k) \lambda(k | A = 1, W) \Big] + S(t | A = 1, W) - \Psi(P)(t)$$

$$\equiv D_{1,t}^*(g_A, S_{A_c}, S) + D_{2,t}^*(P),$$

where

$$h_t(g_A, S_{A_c}, S)(k, A, W) =$$

$$- \frac{I(A = 1) I(k \leqslant t)}{g_A(A = 1 | W) S_{A_c}(k\_ | A, W)} \frac{S(t | A, W)}{S(k | A, W)}.$$

From local least favorable submodel to universal least favorable submodel

- A local least favorable submodel (LLFM) for $S_d(t)$ around initial estimator of conditional hazard:

$$\text{logit}(\lambda_{n,\varepsilon}(\cdot|A=1,W)) = \text{logit}(\lambda_n(\cdot|A=1,W)) + \varepsilon h_t.$$

- Similarly, we have this local least favorable submodel for a vector $(S_d(t) : t)$ by adding vector $(h_t : t)$ extension.
- These imply, as above, universal least favorable submodels for single and multidimensional survival function.

Simulations for one-step TMLE of survival curve

We investigated the performance of one-step TMLE for treatment specific survival curve in two simulation settings.

Data structure

- $O = (W, A, T) \sim P_0$
- $A \in \{0, 1\}$
- treatment intervention: $W \rightarrow d(W) = 1$
- $S_d(t)$ is defined by

$$\Psi(P)(t) = E_P \left[P \left(T > t | A = d(W), W \right) \right]$$

Candidate estimators

1. **Kaplan Meier)**
2. **Iterative TMLE for each single t separately**
3. **One-step TMLE targeting the whole survival curve S_d**

Results

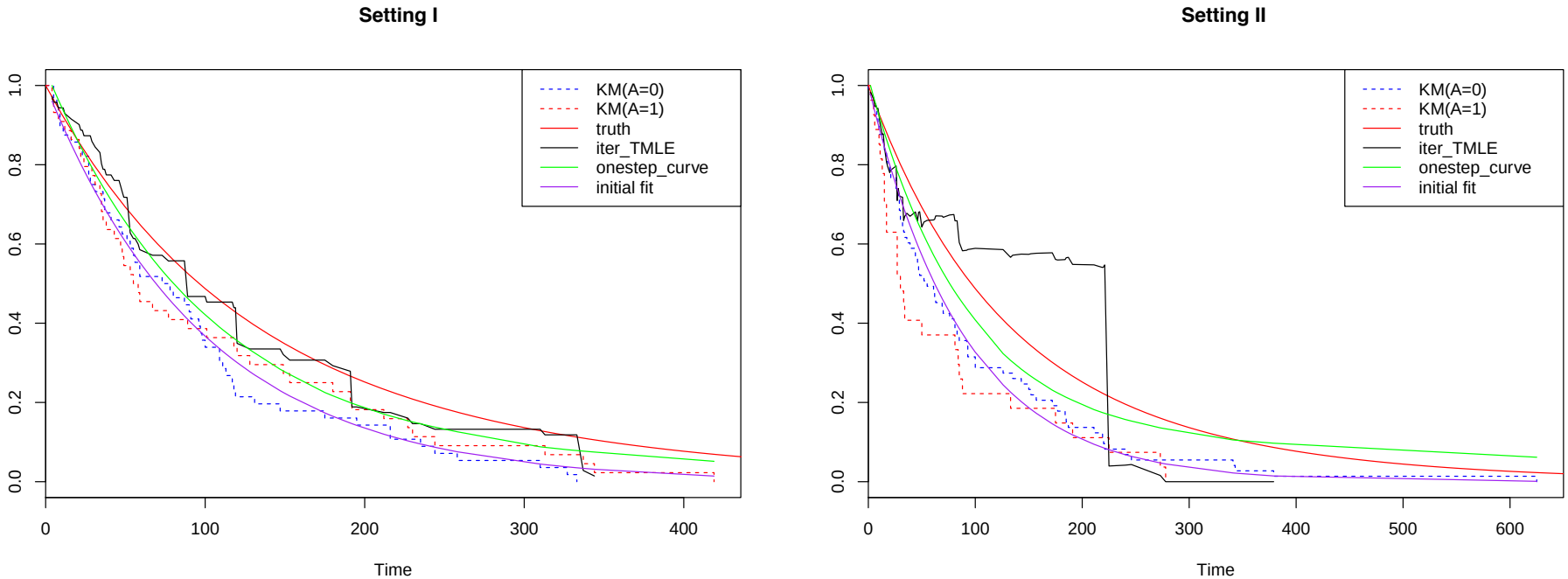


Figure: Based on one data set

Monte-carlo results ($n = 100$)

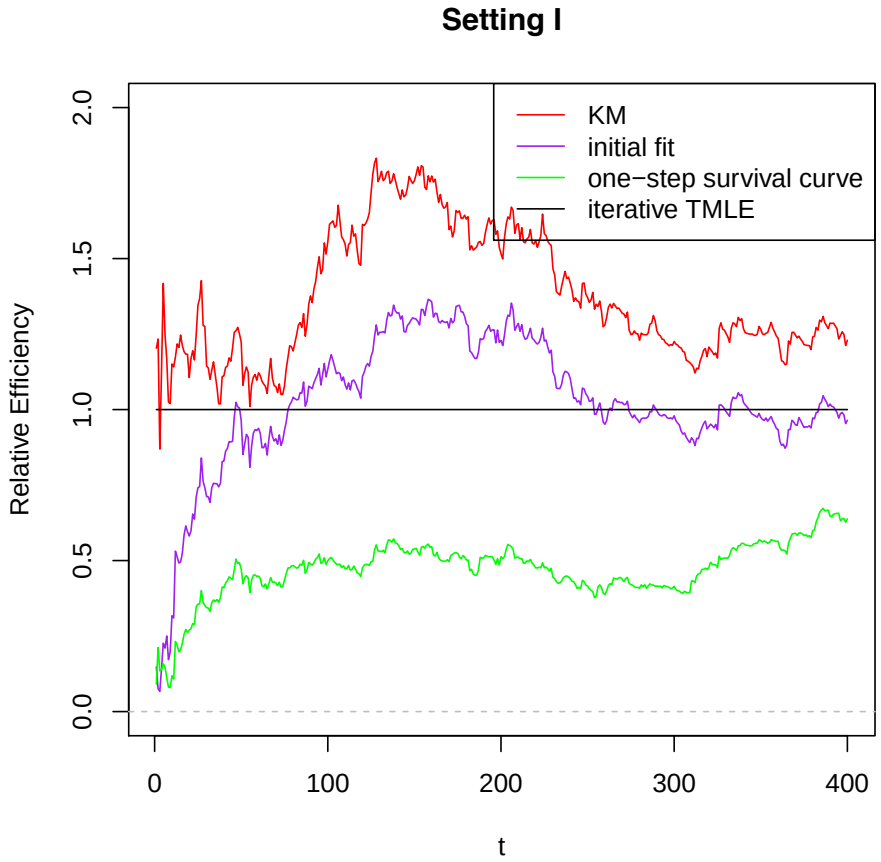


Figure: Relative efficiency against iterative TMLE, as a function of t

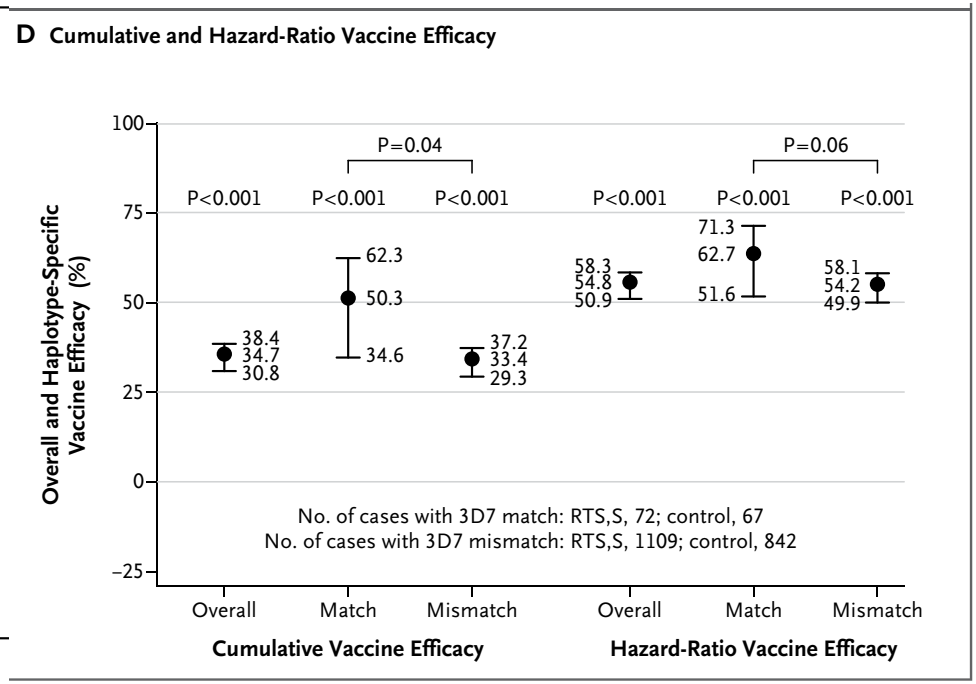
Estimating the impact of genetic polymorphisms on the efficacy of malaria vaccine on the time to infection

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine

D.E. Neafsey, M. Juraska, T. Bedford, D. Benkeser, C. Valim, A. Griggs, M. Lievens, S. Abdulla, S. Adjei, T. Agbenyega, S.T. Agnandji, P. Aide, S. Anderson, D. Ansong, J.J. Aponte, K.P. Asante, P. Bejon, A.J. Birkett, M. Bruls, K.M. Connolly, U. D'Alessandro, C. Dobaño, S. Gesase, B. Greenwood, J. Grimsby, H. Tinto, M.J. Hamel, I. Hoffman, P. Kamthunzi, S. Kariuki, P.G. Kremsner, A. Leach, B. Lell, N.J. Lennon, J. Lusingu, K. Marsh, F. Martinson, J.T. Molel, E.L. Moss, P. Njuguna, C.F. Ockenhouse, B. Ragama Ogutu, W. Otieno, L. Otieno, K. Otieno, S. Owusu-Agyei, D.J. Park, K. Pellé, D. Robbins, C. Russ, E.M. Ryan, J. Sacarlal, B. Sogoloff, H. Sorgho, M. Tanner, T. Theander, I. Valea, S.K. Volkman, Q. Yu, D. Lapierre, B.W. Birren, P.B. Gilbert, and D.F. Wirth



Outline

General Longitudinal Data Structure

We observe n i.i.d. copies of a longitudinal data structure

$$O = (L(0), A(0), \ldots, L(K), A(K), Y = L(K+1)),$$

where $A(t)$ denotes a discrete valued intervention node, $L(t)$ is an intermediate covariate realized after $A(t-1)$ and before $A(t)$, $t = 0, \ldots, K$, and Y is a final outcome of interest.

Survival example: For example,

$$\begin{aligned}
A(t) &= (A_1(t), A_2(t)) \\
A_1(t) &= I(\text{Treated at time } t) \\
A_2(t) &= I(\min(T, C) \leq t, \Delta = 0) \text{ right-censoring indicator process} \\
\Delta &= I(T \leq C) \text{ failure indicator} \\
Y(t) &= I(\min(T, C) \leq t, \Delta = 1) \text{ survival indicator process} \\
Y(t) &\subset L(t) \; Y = Y(K+1).
\end{aligned}$$

Likelihood and Statistical Model

The probability distribution P_0 of O can be factorized according to the time-ordering as

$$\begin{aligned} p_0(O) &= \prod_{t=0}^{K+1} p_0(L(t) \mid Pa(L(t))) \prod_{t=0}^{K} p_0(A(t) \mid Pa(A(t))) \\ &\equiv \prod_{t=0}^{K+1} q_{0,L(t)}(O) \prod_{t=0}^{K} g_{0,A(t)}(O) \\ &\equiv q_0 g_0, \end{aligned}$$

where $Pa(L(t)) \equiv (\bar{L}(t-1), \bar{A}(t-1))$ and $Pa(A(t)) \equiv (\bar{L}(t), \bar{A}(t-1))$ denote the parents of $L(t)$ and $A(t)$ in the time-ordered sequence, respectively. The g_0-factor represents the intervention mechanism.
Statistical Model: We make no assumptions on q_0, but could make assumptions on g_0.

Statistical Target Parameter: G-computation Formula for Post-dynamic-Intervention Distribution

- $p_0^{g^*} = q_0(o)g^*(o)$ is the G-computation formula for the post-intervention distribution of O under the stochastic intervention $g^* = \prod_{t=0}^{K} g^*_{A(t)}(O)$.
- In particular, for a dynamic intervention $d = (d_t : t = 0, \ldots, K)$ with $d_t(\bar{L}(t), \bar{A}(t-1))$ being the treatment at time t, the G-computation formula is given by

$$p_0^d(l) = \prod_{t=0}^{K+1} q^d_{0,L(t)}(\bar{l}(t)),$$

where $q^d_{L(t)}(\bar{l}(t)) = q_{L(t)}(l(t) \mid \bar{l}(t-1), \bar{A}(t-1) = \bar{d}_{t-1}(\bar{l}(t-1)))$.

- Let $L^d = (L(0), L^d(1), \ldots, Y^d = L^d(K+1))$ denote the random variable with probability distribution P^d.

Outline

A Sequential Regression G-computation Formula (Bang, Robins, 2005)

- By the iterative conditional expectation rule (tower rule), we have

$$E_{P^d} Y^d = E \ldots E(E(Y^d \mid \bar{L}^d(K)) \mid L^d(K-1)) \ldots \mid L(0)).$$

- In addition, the conditional expectation, given $\bar{L}^d(K)$ is equivalent with conditioning on $\bar{L}(K), \bar{A}(K-1) = \bar{d}_{K-1}(\bar{L}(K-1))$.

In this manner, one can represent $E_{P^d} Y^d$ as an iterative conditional expectation, first take conditional expectation, given $\bar{L}^d(K)$ (equivalent with $\bar{L}(K), \bar{A}(K-1)$), then take the conditional expectation, given $\bar{L}^d(K-1)$ (equivalent with $\bar{L}(K-1), \bar{A}(K-2)$), and so on, until the conditional expectation given $L(0)$, and finally take the mean over $L(0)$.

TMLE

- A likelihood based TMLE was developed (van der Laan, Stitelman, 2010).
- A sequential regression TMLE $\Psi(Q_n^*)$ was developed for EY_d in van der Laan, Gruber (2012).
- The latter builds on Bang and Robins (2005) by putting their innovative double robust efficient estimating equation method, which uses sequential clever covariate regressions to estimate the nuisance parameters of estimating equation, into aTMLE framework.
- A TMLE for Euclidean summary measures of $(EY_d : d \in \mathcal{D})$ defined by marginal structural working models is developed in Petersen et al. (2013);
- A new (analogue to sequential regression) TMLE allowing for *continuous valued monitoring and time till event* is coming (Rijtgaard, van der Laan, 2019).

Outline

`ltmle` package (Petersen et al. (2013), van der Laan, Gruber (2012), Schitzer et al. (2013))

`R` package: `ltmle`

- Causal effect estimation with multiple intervention nodes
  - Intervention-specific mean under longitudinal dynamic interventions
  - Dynamic marginal structural working models
- General longitudinal data structures
  - Repeated measures outcomes, survival
  - Right censoring
  - Inference: different variance estimators.
- Estimators
  - IPTW
  - Non-targeted MLE
  - TMLE (two algorithms for MSM)
- Options include nuisance parameter estimation via glm regression formulas or calling `SuperLearner()`

Outline

A real-world CER study comparing different rules for treatment intensification for diabetes

- Data extracted from diabetes registries of 7 HMO research network sites:
  - Kaiser Permanente
  - Group Health Cooperative
  - HealthPartners
- Enrollment period: Jan 1^{st} 2001 to Jun 30^{th} 2009

Enrollment criteria:

- past A1c$<$ 7% (glucose level) while on 2+ oral agents or basal insulin
- 7% $\leq$ latest A1c $\leq$ 8.5% (study entry when glycemia was no longer reined in)

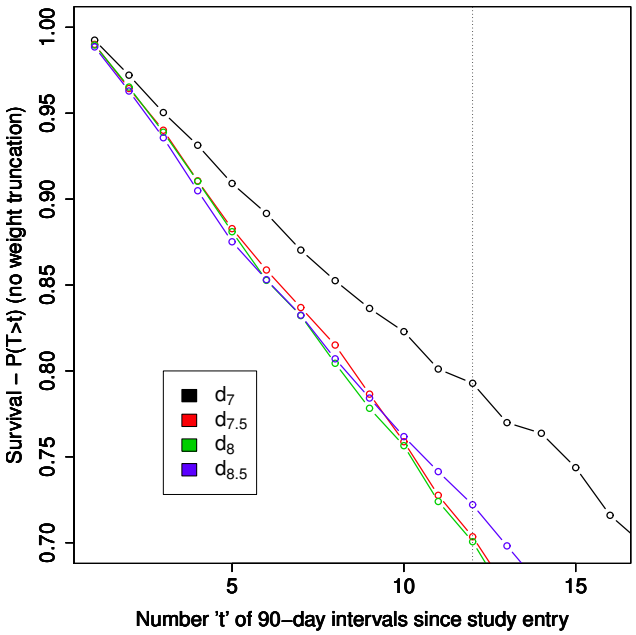
Longitudinal data

- Follow-up til the earliest of Jun 30^{th} 2010, death, health plan disenrollment, or the failure date
- Failure defined as onset/progression of albuminuria (a microvascular complication)
- Treatment is the indicator being on "treatment intensification" (TI)
- $n \approx 51,000$ with a median follow-up of 2.5 years

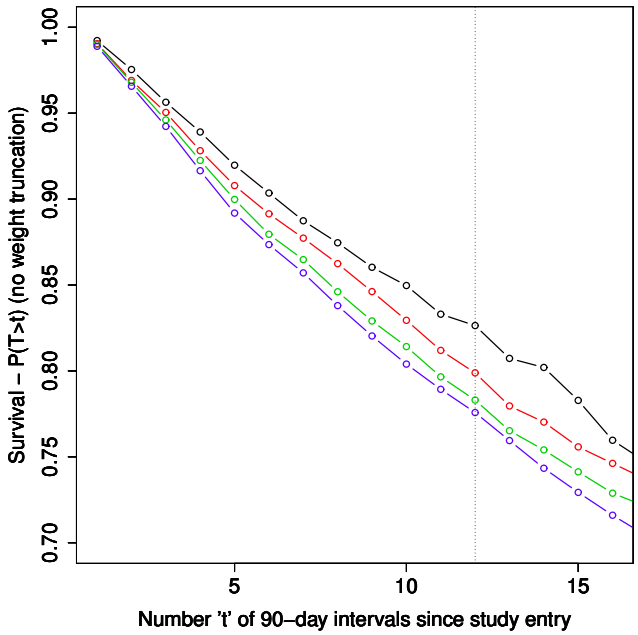
Back to the TI study...

Impact of machine learning on inference with IPW 1:

Parametric model



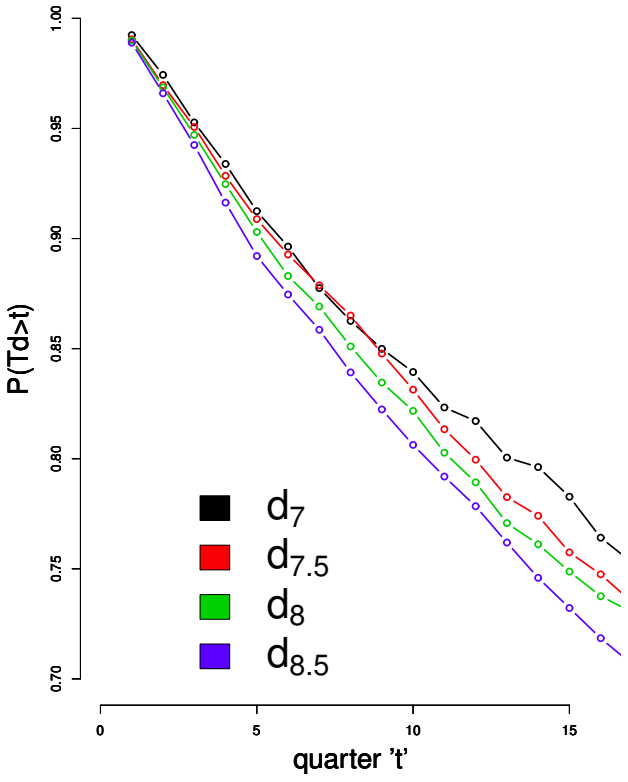
Super Learning



No/weak evidence of protective effect

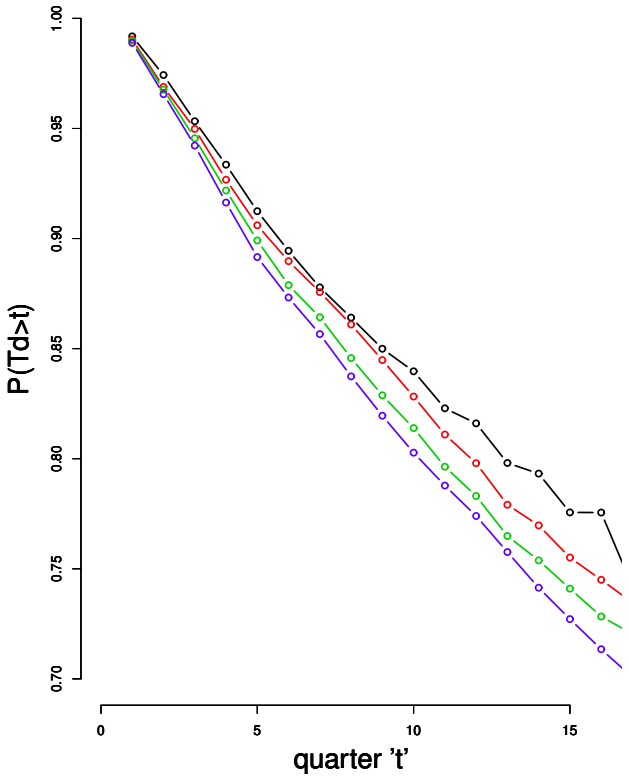
Strong significant evidence

Practical performance



IPW estimator 3 + SL

(hazard-based)



TMLE + SL

$$1.07 \leq \frac{\sigma_{IPW3}}{\sigma_{TMLE}} \leq 1.11$$

Analysis of Tshepo Study (Gruttolas et al.): RCT with time-to-event outcome

TMLE of causal effect of point treatment on survival needs to account for potential bias due to informative censoring by **time-dependent covariates** CD4 and viral load that have an effect on both time to drop-out and the time to event of interest.

We will directly compare results using this TMLE that only incorporates the baseline covariates to the TMLE that accounts for time dependent confounding in the form of informative censoring due to the time-dependent covariates. Moreover, we will compare these results to results based on an **IPCW estimator** and a locally efficient double robust **estimating equation based estimator** (A-IPCW).

Analysis of Tshepo Study (Gruttolas et al.): RCT with time-to-event outcome

For the analysis performed here we evaluate the **effect modification by gender** on the two cART treatments for Time to death censored by treatment modification or end of study (DEATH).
We estimate the difference in additive risk by gender at 36 months after randomization to cART therapy. We will estimate this effect modification parameter using the six estimators.

Results of Tshepo Study (Gruttolas et al.)

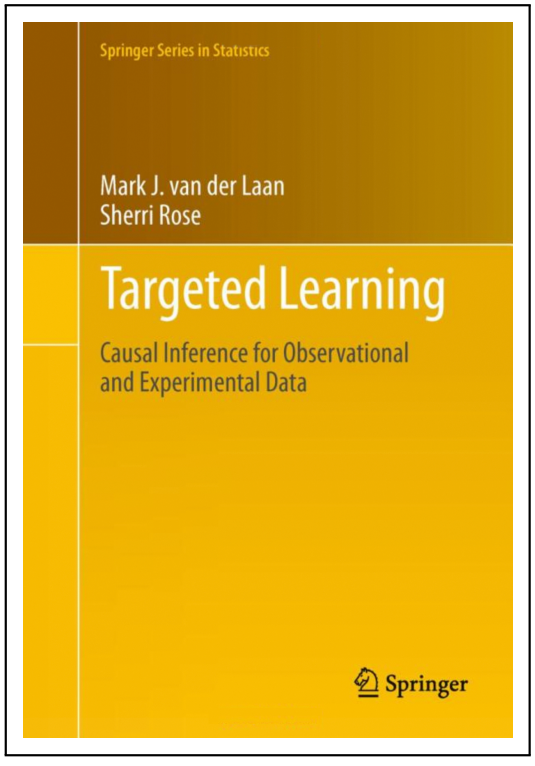
Risk Difference @ 36 Months

	Time Dependent			Baseline		
	TMLE	A-IPCW	IPCW	TMLE	A-IPCW	IPCW
Est	6.3%	6.5%	5.2%	5.1%	5.1%	5.2%
SE	2.3%	2.3%	12.5%	2.4%	2.4%	12.5%
p-value	0.005	0.004	0.680	0.029	0.030	0.680

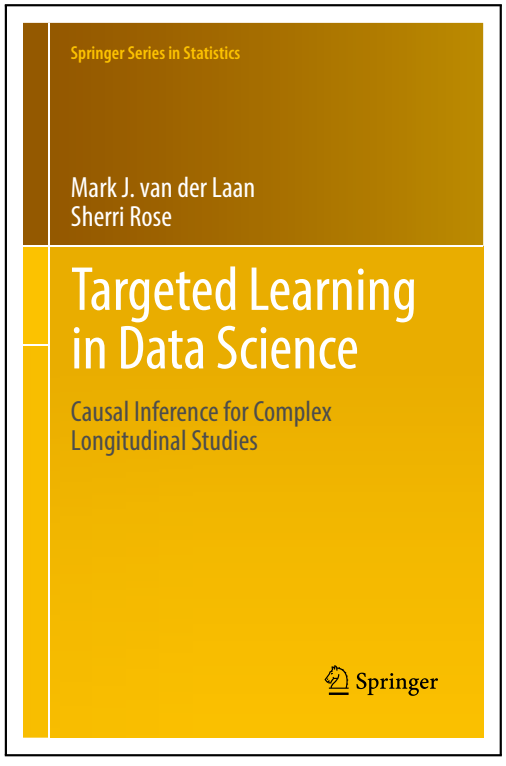
Table: Gender Effect Modification on Death

The TMLE results indicate that gender does in fact modify the effect of the drug treatment EFV/NVP and the difference in the effect between males and females at 36 months is estimated at 6.3 percent.

Targeted Learning



van der Laan & Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer, 2011.



van der Laan & Rose, *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. New York: Springer, 2018.

`https://vanderlaan-lab.org`

Mark's Blog

- You're welcome to submit a question to my blog by sending an email to `vanderlaan.blog@berkeley.edu`.
- I answer all questions myself, keeping the inquirer anonymous.
- Past posts: `https://vanderlaan-lab.org/post/`