Nonparametric doubly-robust inference on the average treatment effect

Marco Carone
Department of Biostatistics
School of Public Health, University of Washington

SCHOOL OF PUBLIC HEALTH
UNIVERSITY *of* WASHINGTON

Joint work with
David Benkeser (Emory U.), **Peter Gilbert** (Fred Hutch)
and **Mark van der Laan** (UC Berkeley)

UW Causal Working Group
April 13, 2018

Background: doubly-robust strategies for ATE estimation

If $A \in \{0, 1\}$ is a binary treatment and $Y(a)$ the counterfactual outcome corresponding to treatment level $A = a$, we are often interested in the average treatment effect

$$ATE \;:=\; E[Y(1) - Y(0)] \;=\; E[Y(1)] - E[Y(0)] \;.$$

The observed data consist of $O_1, O_2, \ldots, O_n \overset{iid}{\sim} P_0$, with $O_i := (W_i, A_i, Y_i)$ and

W_i = the vector of baseline patient characteristics (i.e., potential confounders);
A_i = the treatment/intervention received;
Y_i = the outcome of interest.

Two fundamental questions:

- When/how can $E[Y(a)]$ be identified (i.e., estimated) from the observed data?
- In such case, what estimation procedure should we use?

Background: doubly-robust strategies for ATE estimation

Key condition #1: **the randomization** (or ignorability) **condition**

Holds if treatment is randomized within strata of recorded covariates:

$$(Y(0), Y(1)) \perp A \mid W$$

This will generally hold if all potential confounders have been recorded.

Key condition #2: **the positivity** (or experimental treatment assignment) **condition**

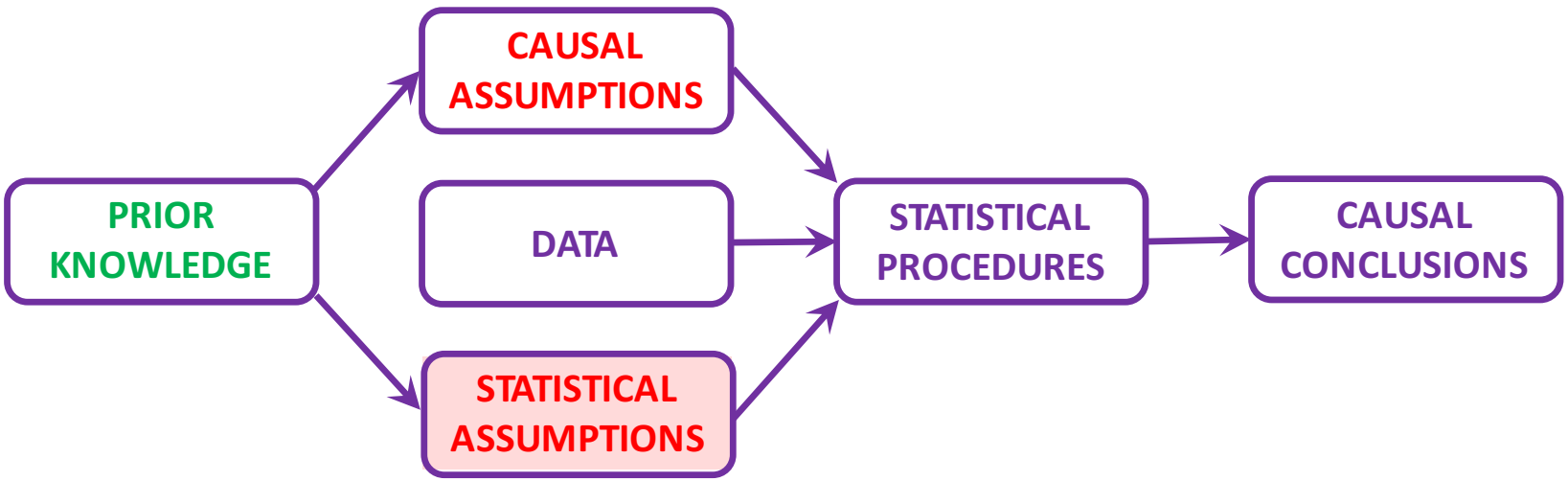
Holds provided each patient may potentially be assigned to any treatment group:

$$\text{for each } a,\ P(A = a \mid W = w) > 0 \text{ for every plausible value } w$$

The **G-computation formula** provides an identification of the causal quantity:

$$E[Y(a)] = E\left[E(Y \mid A = a, W)\right] .$$

Background: doubly-robust strategies for ATE estimation



Goal of this work: provide valid inference with few statistical assumptions!

Background: doubly-robust strategies for ATE estimation

We will focus on the G-computation statistical parameter corresponding to $A = 1$:

$$\Psi(P) := E_P\left[E_P\left(Y \mid A = 1, W\right)\right].$$

Two observed data quantities play a key role in nearly all methods for causal inference:

$$\begin{aligned} \textbf{the outcome regression :} \quad & \bar{Q}_P(w) := E_P(Y \mid A = 1, W = w) \\ \textbf{the propensity score :} \quad & g_P(w) := P(A = 1 \mid W = w) \quad . \end{aligned}$$

Estimators of $\psi_0 := \Psi(P_0)$ build upon estimators of $\bar{Q}_0 := \bar{Q}_{P_0}$ and/or $g_0 := g_{P_0}$.

In the following, we will denote by $\bar{Q}_n$ and g_n estimators of $\bar{Q}_0$ and g_0, respectively.

Background: doubly-robust strategies for ATE estimation

As estimator of ψ_0, we could simply use the G-estimator based on $\bar{Q}_n$:

$$\psi_{n,G} := \frac{1}{n} \sum_{i=1}^{n} \bar{Q}_n(W_i) \ .$$

If $\bar{Q}_n$ is a *consistent* parametric estimator of $\bar{Q}_0$, life is good!

Usually, we cannot correctly specify a parametric model a priori. This motivates us to use of flexible, data-driven estimation strategies. Then, life becomes complicated...

The problem comes from the fact that $\Psi(\bar{Q}_n)$ is generally too biased. However, this bias can be approximated by the computable quantity

$$B_n(\bar{Q}_n, g_n) := -\frac{1}{n} \sum_{i=1}^{n} \frac{A_i}{g_n(W_i)} \left[Y_i - \bar{Q}_n(W_i) \right] \ .$$

Background: doubly-robust strategies for ATE estimation

This readily suggests two strategies for constructing improved estimators of ψ_0:

- the AIPTW estimator of Robins et al. (1994):

$$\psi_{n,AIPTW} := \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n(W_i) - B_n(Q_n, g_n)$$

- the TMLE estimator of van der Laan & Rubin (2006):

$$\psi_{n,TMLE} := \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^*(W_i) \text{ with } \bar{Q}_n^* \text{ such that } B_n(\bar{Q}_n^*, g_n) = 0$$

These estimators have the following properties:

1. **Local efficiency:** If $\bar{Q}_n$ and g_n are both consistent, they are efficient and can be used to perform valid inference.
2. **Double robustness:** If at least one of $\bar{Q}_n$ and g_n is consistent, they are consistent.

Achieving valid doubly-robust inference

The AIPTW and TMLE estimators discussed so far are said to be doubly-robust. To be precise, we should say that they enjoy doubly-robust consistency.

When both $\bar{Q}_n$ and g_n are consistent, we can easily construct valid CIs and p-values.

What about when only one of $\bar{Q}_n$ and g_n is consistent?

- If parametric models are used, the bootstrap can be used. Improved procedures have also been proposed for this case (e.g., Vermeulen & Vansteelandt, 2015).
- If flexible estimation techniques are used, we are out of luck – there is generally no nice limit distribution we can lean on.

SCENARIO I: $\bar{Q}_n \xrightarrow{P} \bar{Q}_0$ but $g_n \xrightarrow{P} g \neq g_0$

Denoting by EIF the nonparametric efficient influence function of Ψ, we have that

$$\psi_n - \psi_0 \approx \frac{1}{n}\sum_{i=1}^{n} EIF(\bar{Q}_0, g)(O_i) + \int \underbrace{[\bar{Q}_n(w) - \bar{Q}_0(w)]}_{\downarrow 0} \underbrace{\left[\frac{g_n(w) - g_0(w)}{g_n(w)}\right]}_{\not\downarrow 0} dQ_{W,0}(w) .$$

Achieving valid doubly-robust inference

Setting $\tilde{\Psi}(\bar{Q}) := \int \bar{Q}(w) \left[\frac{g(w)-g_0(w)}{g(w)}\right] dQ_{W,0}(w)$, with some work, we can write

$$\begin{aligned}
\text{remainder term} &= \int \left[\bar{Q}_n(w) - \bar{Q}_0(w)\right] \left[\frac{g(w) - g_0(w)}{g(w)}\right] dQ_{W,0}(w) \\
&= \tilde{\Psi}(\bar{Q}_n) - \tilde{\Psi}(\bar{Q}_0) \\
&\approx \frac{1}{n}\sum_{i=1}^{n} A_i \frac{g_{2,0,r}(W_i)}{g_{1,0,r}(W_i)} \left[Y_i - \bar{Q}_0(W_i)\right] + B_{I,n}(g_{1,n,r}, g_{2,n,r}, \bar{Q}_n) \ ,
\end{aligned}$$

where $g_{1,n,r}$ and $g_{2,n,r}$ are consistent estimators of

$$\begin{aligned}
g_{1,0,r}(w) &:= E[A \mid \bar{Q}_0(W) = \bar{Q}_0(w)] \\
g_{2,0,r}(w) &:= E[\{A - g(W)\}/g(W) \mid \bar{Q}_0(W) = \bar{Q}_0(w)] \ ,
\end{aligned}$$

and inconsistent estimation of g_0 yields the additional bias term

$$B_{I,n}(g_{1,n,r}, g_{2,n,r}, \bar{Q}_n) := \frac{1}{n}\sum_{i=1}^{n} A_i \frac{g_{2,n,r}(W_i)}{g_{1,n,r}(W_i)} \left[Y_i - \bar{Q}_n(W_i)\right].$$

Achieving valid doubly-robust inference

We can proceed similarly in **SCENARIO II:** $g_n \xrightarrow{P} g_0$ but $\bar{Q}_n \xrightarrow{P} \bar{Q} \neq \bar{Q}_0$.

Specifically, we find that

$$\text{remainder term} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{Q}_{0,r}(W_i)}{g_0(W_i)}\left[Y_i - \bar{Q}(W_i)\right] + B_{II,n}(\bar{Q}_{n,r}, \bar{Q}_n, g_n)$$

where $\bar{Q}_{n,r}$ is a consistent estimator of

$$\bar{Q}_{0,r}(w) := E[Y - \bar{Q}(W) \mid A = 1, g_0(W) = g_0(w)]$$

and inconsistent estimation of $\bar{Q}_0$ yields the additional bias term

$$B_{II,n}(\bar{Q}_{n,r}, \bar{Q}_n, g_n) := \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{Q}_{n,r}(W_i)}{g_n(W_i)}\left[A_i - g_n(W_i)\right].$$

Achieving valid doubly-robust inference

From initial estimators $\bar{Q}_n$ and g_n, the TMLE framework can be used to find revised estimators $\bar{Q}_n^*$ and g_n^* that not only satisfy $B_n(Q_n^*, g_n^*) = 0$ but also

$$B_{I,n}(g_{1,n,r}^*, g_{2,n,r}^*, \bar{Q}_n^*) = 0 \quad \text{and} \quad B_{II,n}(\bar{Q}_{n,r}^*, \bar{Q}_n^*, g_n^*) = 0 \ ,$$

where $g_{1,n,r}^*$, $g_{2,n,r}^*$ and $\bar{Q}_{n,r}^*$ are estimators based upon $\bar{Q}_n^*$ and g_n^*.

The resulting TMLE estimator is iteratively-defined and has the following properties:

1. is efficient when both $\bar{Q}_n$ and g_n are consistent;
2. is consistent when at least one of $\bar{Q}_n$ and g_n is consistent;
3. when suitably normalized, tends to a mean-zero normal distribution with variance we can consistently estimate, even when only one of $\bar{Q}_n$ or g_n is consistent.

It seems that the AIPTW estimator cannot be adapted to yield valid DR inference.

Details are provided in **Benkeser, Carone, van der Laan & Gilbert (2017)**, and build upon ideas in **van der Laan (2014)**.

Achieving valid doubly-robust inference

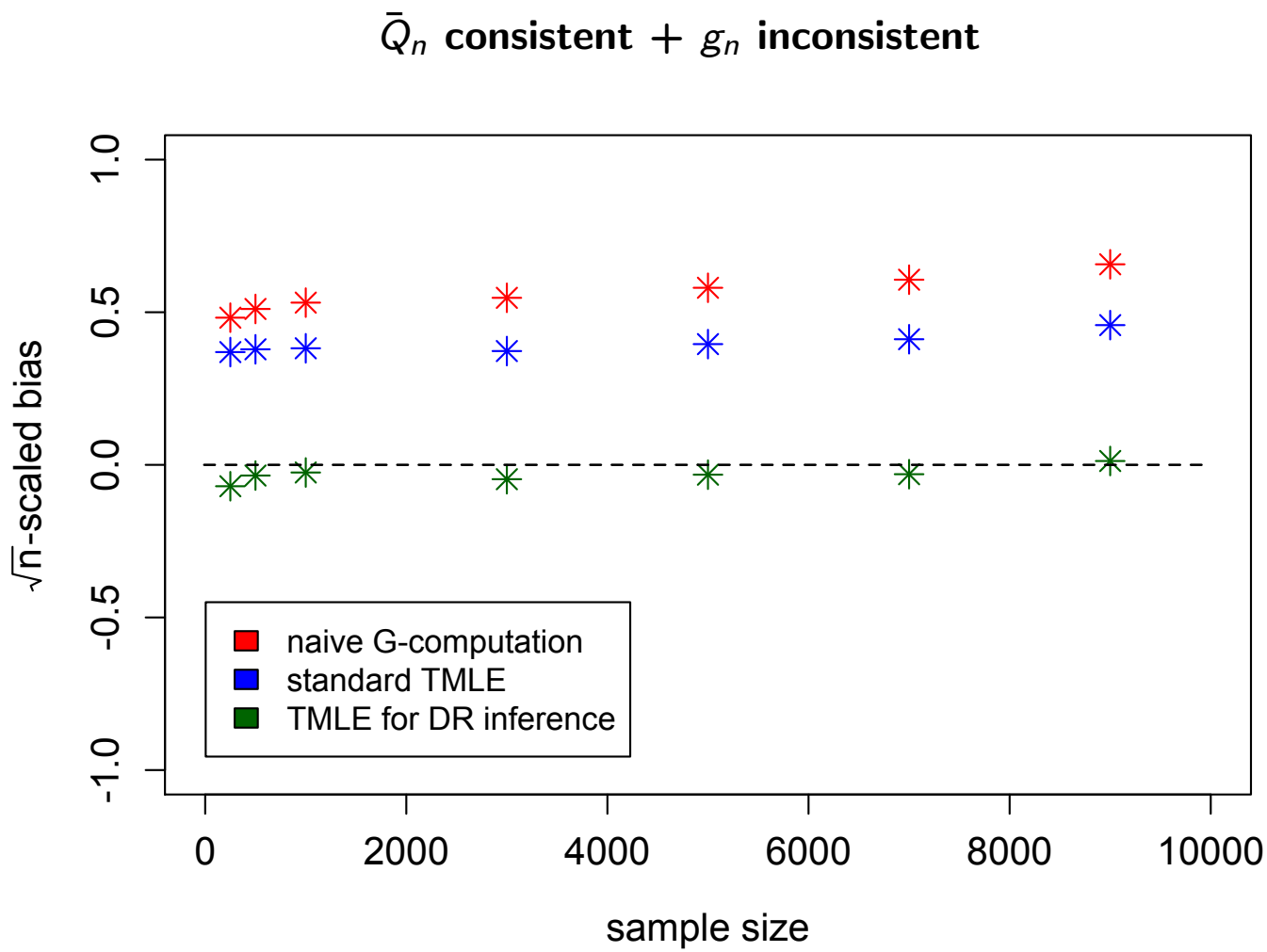
In a simulation study, data were generated as follows:

1. $W = (W_1, W_2)$ with $W_1 \sim U(-2, +2)$, $W_2 \sim \text{Bernoulli}(0.5)$ and $W_1 \perp W_2$
2. $A \mid W = w \sim \text{Bernoulli}(g_0(w))$ with $g_0(w) := \text{expit}(-w_1 + 2w_1w_2)$
3. $Y \mid A = a, W = w \sim \text{Bernoulli}(\bar{Q}_0(a, w))$ with $\bar{Q}_0(a, w) := \text{expit}(0.2a - w_1 + 2w_1w_2)$

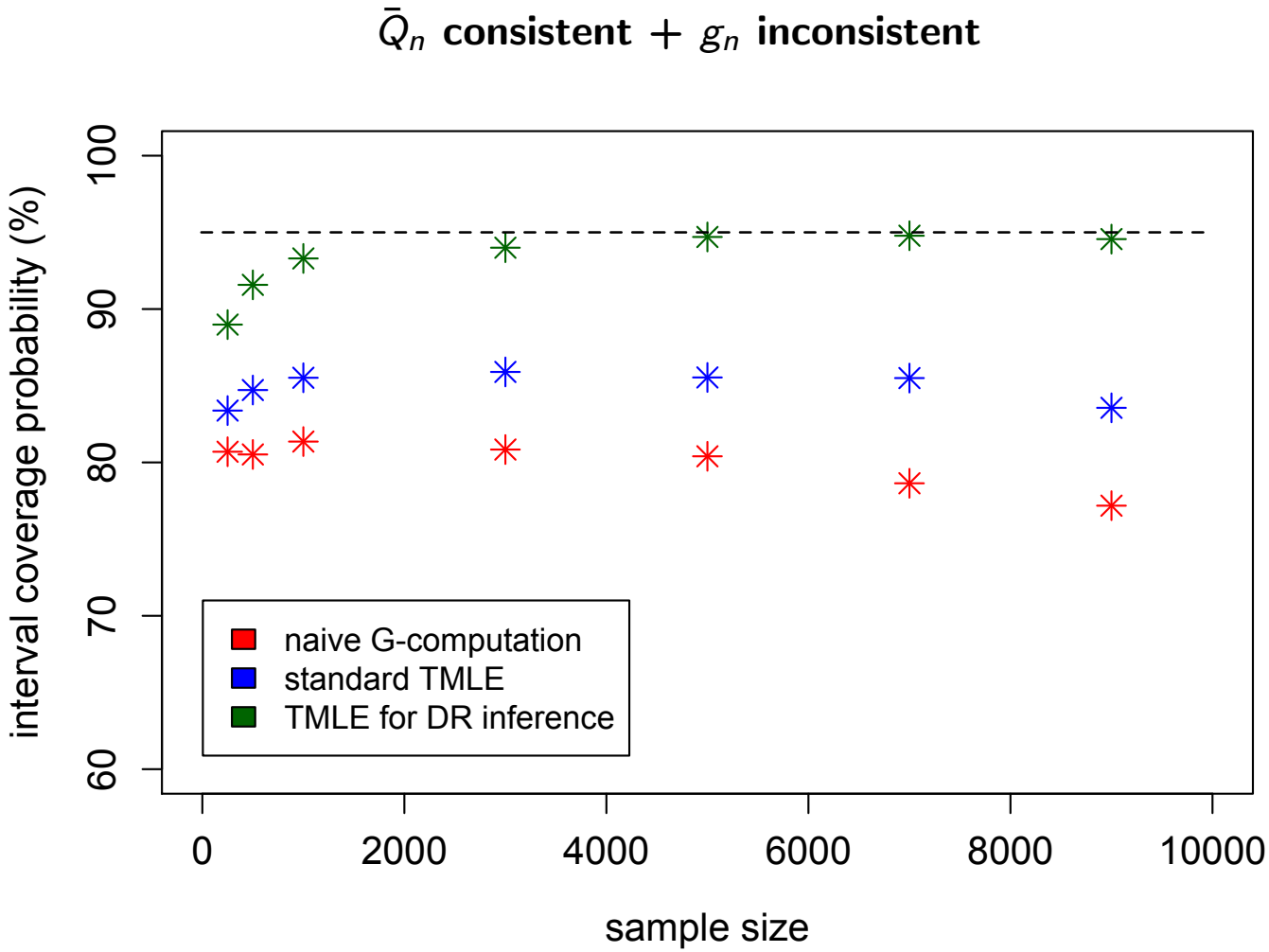
In different settings, $\bar{Q}_0$ and $\bar{g}_0$ were estimated either

- **consistently:** with a bivariate kernel smoother with CV-selected bandwidth;
- **inconsistently:** with a main terms-only logistic regression.

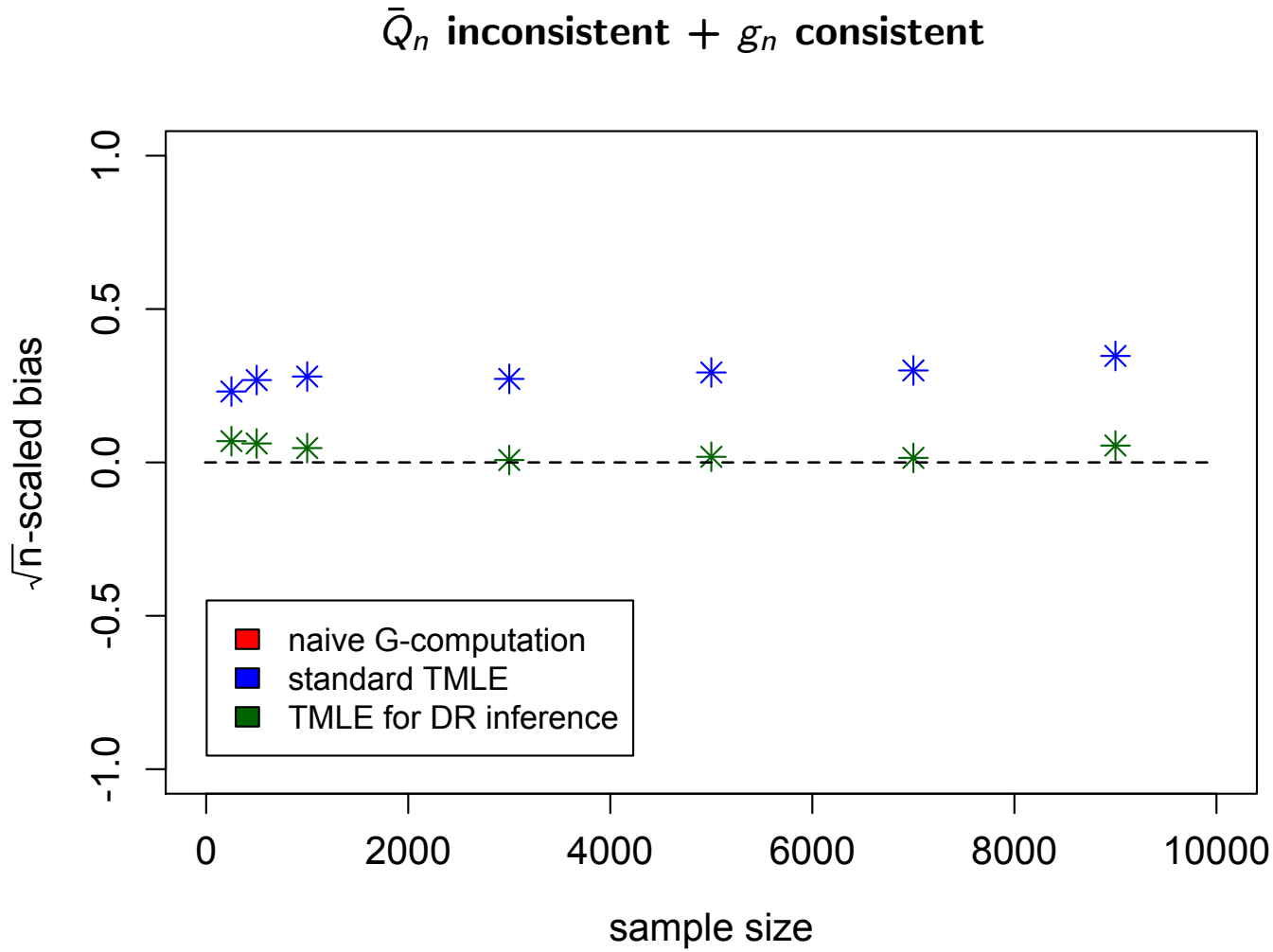
Achieving valid doubly-robust inference



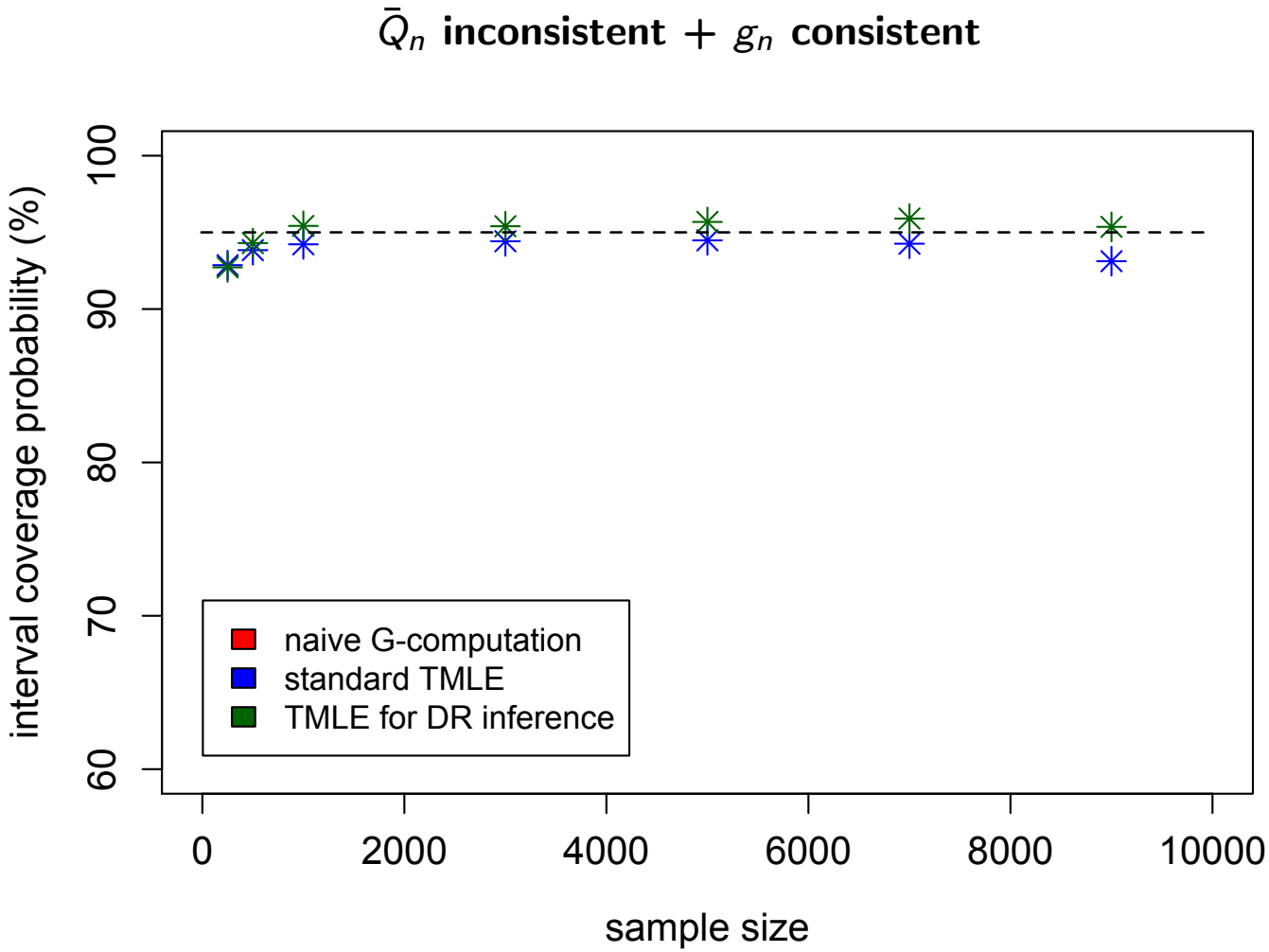
Achieving valid doubly-robust inference



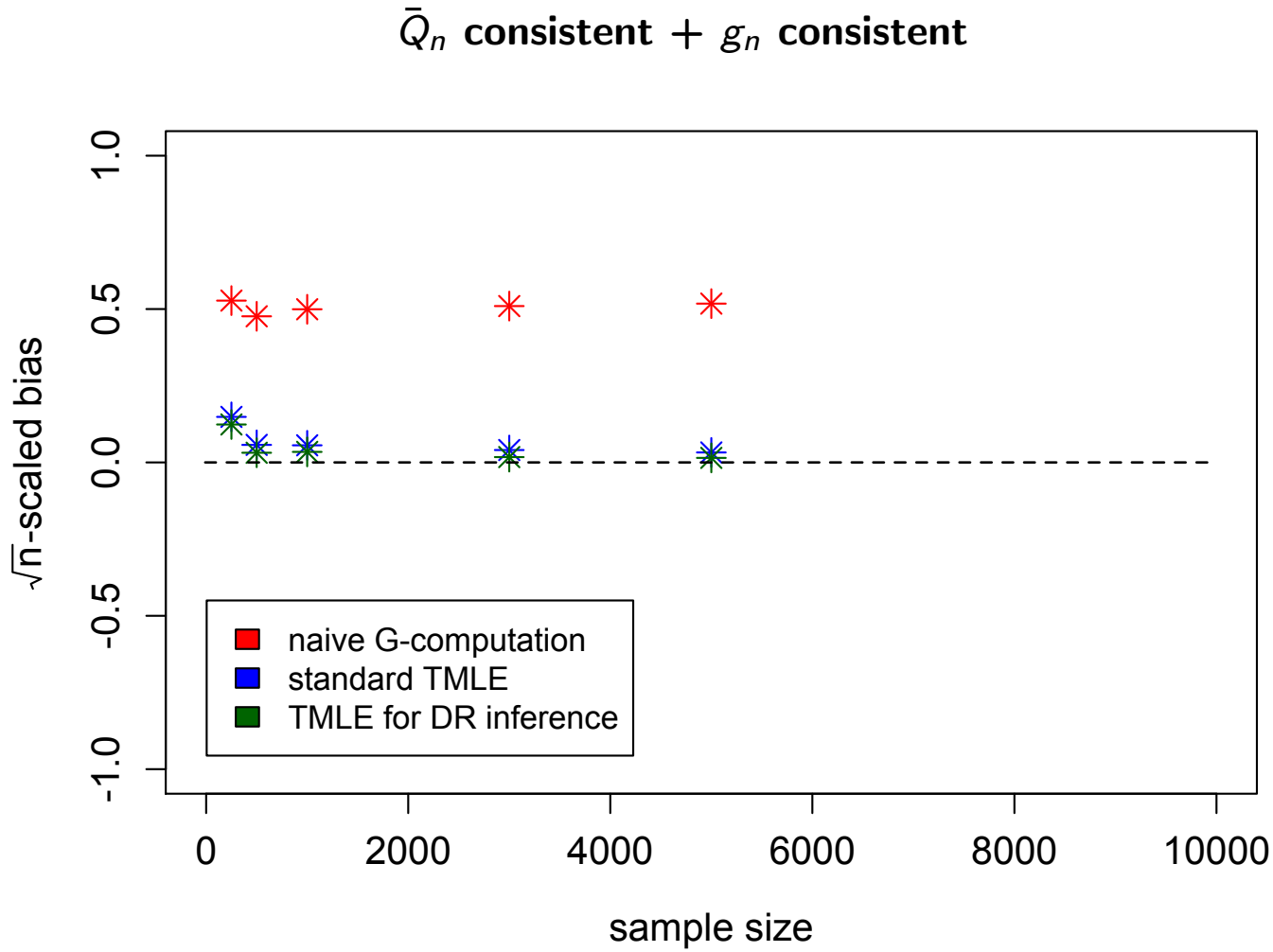
Achieving valid doubly-robust inference



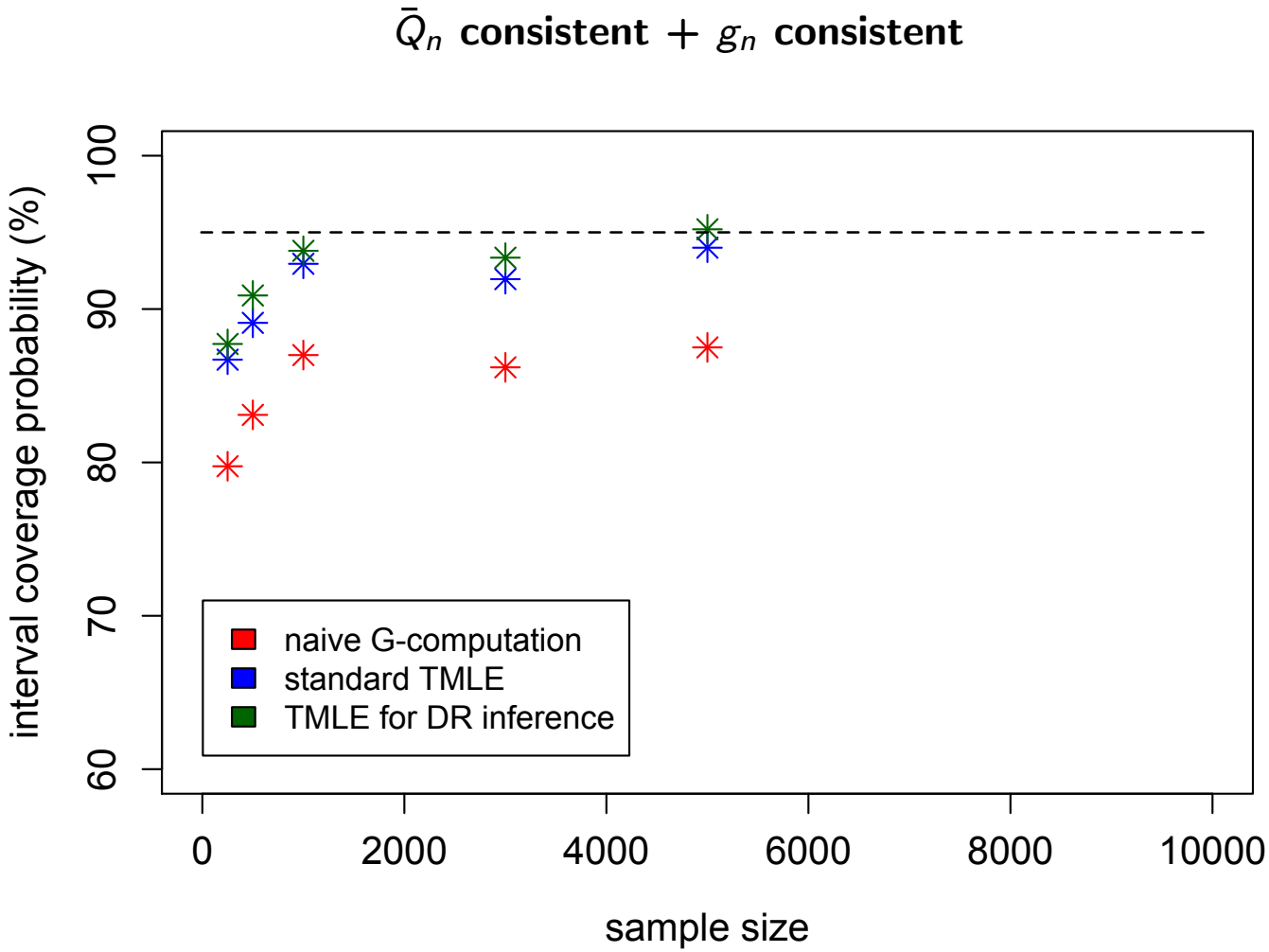
Achieving valid doubly-robust inference



Achieving valid doubly-robust inference



Achieving valid doubly-robust inference



Achieving valid doubly-robust inference

BACK PAIN OUTCOMES USING LONGITUDINAL DATA (BOLD) STUDY

- observational study of 5239 patients of age $\geq$ 65 years with new primary care visit for back pain between 2010 and 2013
- three integrated systems:
  - Kaiser Permanente (Northern CA)
  - Henry Ford Health System (Detroit, MI)
  - Harvard Vanguard / Harvard Pilgrim (Boston, MA)
- patients identified through Health Care Information Systems
- contacted at 3, 6, 12 and 24 months
- main outcomes: pain, disability (Roland Morris Disability Questionnaire – RMDQ), depression and anxiety

A key scientific question investigators wished to address:

In seniors with a new visit for back pain, how effective is early imaging compared with no imaging in reducing the RMDQ score at 12 months?

Achieving valid doubly-robust inference

We analyzed data from the BOLD study using TMLE and estimation of the propensity score and outcome regression using Super Learner ensembling (including GLM, generalized additive models and Lasso regression).

We estimated the ATE comparing early imaging to control.

without correction for DR inference:

$$\text{estimate} = \text{-0.36}, \quad 95\% \text{ CI: } (\text{-0.72}, \text{-0.00}), \quad p = 0.05$$

with correction for DR inference:

$$\text{estimate} = \text{-0.43}, \quad 95\% \text{ CI: } (\text{-0.77}, \text{-0.09}), \quad p = 0.02$$

Based on these results, we would conclude that obtaining early imaging appears to lower disability scores on average at the 12-month mark.

What about randomized trials?

In observational studies, adjustment for confounding is mandatory. However, even in randomized trials, the methods discussed can be useful and sometimes needed.

In an ideal randomized trial, use of G-computation is not necessary since then

$$ATE = E(Y \mid A = 1) - E(Y \mid A = 0)$$

and the unadjusted difference-in-means estimator is consistent.

We could still use the TMLE estimator with g_n = true and known propensity score:

- ⊕ Because g is exactly know, consistency is guaranteed by double robustness.
- ⊕ Valid confidence intervals and p-values can be obtained even if $\bar{Q}_n$ is inconsistent.
- ⊕ If baseline covariates are predictive of the outcome, their inclusion generally yields tighter confidence intervals / more powerful tests.

See Moore & van der Laan (2009) for more details.

What about randomized trials?

In most trials, some aspects of the study are beyond the control of the investigators.

For example, the outcome Y may be missing for some trial participants not by design. To ensure consistency, we may want to adjust for available covariates as best we can.

Write $\Delta := I(\text{outcome not missing})$. The data unit is then $O := (W, A, \Delta, \Delta Y)$.

If Y is missing at random given treatment assignment A and covariate vector W,

$$E\left[Y(1)\right] = E\left[E\left(\tilde{Y} \mid \tilde{A} = 1, W\right)\right]$$

for $\tilde{Y} := \Delta Y$ and $\tilde{A} := \Delta A$. The proposed TMLE can be used to draw DR inference.

What about randomized trials?

Loss to follow-up is frequent in trials, especially when focus is on a time-to-event.

Common approaches to deal with loss to follow-up in a trial:

1. **contrast KM curves;**
   - ⊕ preserve interpretation of estimand as a population-averaged causal effect;
   - ⊖ requires independent censoring mechanism in each group!
2. **fit Cox model and report hazard ratio estimate.**
   - ⊕ only requires independent censoring within covariate and treatment strata;
   - ⊖ does not readily provide an interpretation as a population-average causal effect;
   - ⊖ if hazards are not proportional, all bets are off: the estimand will not generally have a causal interpretation.

We could reframe the problem in terms of a longitudinal multi-point treatment study.

Extension to multi-timepoint settings

From a longitudinal multi-timepoint treatment study, we observe data of the form

$$\underbrace{L_0 \rightarrow A_0}_{t_0} \rightarrow \underbrace{L_1 \rightarrow A_1}_{t_1} \rightarrow \cdots \rightarrow \underbrace{L_K \rightarrow A_K}_{t_K} \rightarrow \underbrace{Y}_{t_{K+1}},$$

where we have defined components

$$\begin{aligned} L_k &= \text{covariates recorded at time } t_k; \\ A_k &= \text{treatment assignment at time } t_k; \\ Y &= \text{outcome recorded at some fixed time } t_{K+1}. \end{aligned}$$

We may be interested in the mean value $E[Y(1, 1, \ldots, 1)]$ of the counterfactual outcome defined by setting all treatment nodes to 1.

The corresponding G-computation is more complex: under certain causal conditions,

$$E[Y(1, 1, \ldots, 1)] = E\left[E\left[\ldots E\left[E\left[Y \middle| \overline{A}_K = 1, \overline{L}_K\right] \middle| \overline{A}_{K-1} = 1, \overline{L}_{K-1}\right] \ldots \middle| L_0\right]\right],$$

where $\overline{A}_k := (A_0, A_1, \ldots, A_k)$ and $\overline{L}_k := (L_0, L_1, \ldots, L_k)$ for any k.

Work on DR inference in this context began in Benkeser (2015) and is ongoing.

Perspective and closing thoughts

- We now have access to ATE estimators that allow doubly-robust inference in the single-timepoint setting using the TMLE framework.
- An extension to the multi-timepoint setting appears promising.
- Need further practical use to evaluate the real-life (i.e., finite sample) operating characteristics of these new procedures, particularly in the longitudinal case.
- Additional robustness (multiple robustness) is possible in the multi-timepoint setting – see Luedtke et al. (2017). Can we perform multiply-robust inference?
- Overall objective = maximize robustness of statistical/scientific findings.

Perspective and closing thoughts

Many thanks for your attention! Feedback and/or questions? **mcarone@uw.edu**.

GitHub package: `/benkeser/drtmle`

References:

- Robins, J.M., Rotnitzky, A., Zhao, L.P. 1994. Estimation of regression coefficients when some regressors are not always observed. *JASA*.
- van der Laan, M.J., Rubin, D. 2006. Targeted maximum likelihood learning. *The International Journal of Biostatistics*.
- Vermeulen, K., Vansteelandt, S. 2015. Bias-reduced doubly robust estimation. *JASA*.
- Benkeser, D., Carone, M., van der Laan, M.J., Gilbert, P.B. 2017. Nonparametric doubly-robust inference on the average treatment effect. *Biometrika*.
- Moore, K.L., van der Laan, M.J. 2009 Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. *The International Journal of Biostatistics*.
- Benkeser, D. 2015. Data-adaptive estimation in longitudinal data structures with applications in vaccine efficacy trials. *PhD dissertation, University of Washington*.
- Luedtke, A.R., Sofrygin, O., van der Laan, M.J., Carone, M. 2017. Sequential double robustness in right-censored longitudinal models. *ArXiv technical report*.
- van der Laan, M.J. 2014. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*.