A cautionary note for plasmode simulation studies in the setting of causal inference

Pamela Shaw

Kaiser Permanente Washington Health Research Institute
Seattle, WA
pamela.a.shaw@kp.org

WNAR Whistler, British Columbia, Canada
June 18, 2025

KAISER PERMANENTE®

Acknowledgments and Disclosures

Acknowledgments

- This is joint work with
  - Mark van der Laan (University of California, Berkeley),
  - Susan Gruber (TL Revolution, LLC),
  - Rishi Desai (Brigham and Women's Hospital, Harvard Medical School)
  - Brian Williamson, Susan Shortreed, Chloe Krakauer, and Jen Nelson (Kaiser Permanente Washington Health Research institute)
- This research is supported in part by Task Order 75F40123F19006 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA) and National Institutes of Health (NIH) grant R01-AI131771.

Disclosure

The contents are those of the authors and do not necessarily represent the official views of, nor endorsement, by FDA/HHS, National Institutes of Health, or the U.S. Government.

Outline

- Introduction
- Two types of bootstrap: Empirical sampling of treatment and generating treatment
- Why is empirical sampling of treatment biased?
- Synthetic data simulation study
- Real data simulation study
- Conclusions

Plasmode Simulation Introduction

- Assume a sample of n i.i.d. observations $(W_i, A_i, Y_i) \sim P_0$.
- Let the statistical estimand be the ATE: $\Psi(P) = E_P\{E_P(Y \mid A=1, W) - E_P(Y \mid A=0, W)\}$.
- We have a given estimator $\hat{\Psi}(P_n)$ such as an IPTW estimator

$$\hat{\Psi}_{IPTW}(P_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i(2A_i - 1)}{g_n(A_i \mid W_i)}$$

  for an estimator $g_n = \hat{g}(P_n)$ of the true treatment mechanism $g_0(a \mid W) = P_0(A = a|W)$.
- **Plasmode Simulation**: We wish to evaluate the statistical performance of such an estimator based on sampling from a data distribution that resembles P_0 in the sense that the observed behavior will be highly reflective of the behavior of estimator under sampling from P_0.
- For notational convenience, let's focus on estimation of $EY_1 = E_P E_P(Y \mid A = 1, W)$.

Plasmode simulation sampling frameworks

Table: Data generating mechanisms for plasmode simulation approaches.

	Sample Treatment	**Generate Treatment**
Covariates	Sample W with replacement	Sample W with replacement
Treatment	Sample $A = a$ along with W	Generate $A^{\#} \sim f_A(W, U_A)$
Outcome	Generate $Y^{\#} \sim f_Y(A, W, U_Y)$	Generate $Y^{\#} \sim f_Y(A^{\#}, W, U_Y)$

Fundamental problem

The positivity assumption required for identifying a causal estimand, $P(A = a|W) > 0$ for all a and observed W in the data, is violated under the Sample Treatment framework.

- Under this plasmode approach, every time W_i is sampled, the associated value for A_i is fixed at some a_i, its value in the original data for subject i; thus, the probability that $A = a_i$ for W_i is 1.
- Estimators relying on outcome regression for consistency (e.g., parametric G-computation, glm) are fully reliant on extrapolation for the treatment/covariate combinations missed by the Sample Treatment algorithm.
- Estimators relying on propensity score estimation (e.g. IPTW) will end up having non-negligible bias in the plasmode samples, even when the propensity score model is correctly specified.
- The Generate Treatment approach avoids this problem.

Bootstrap Approach 1: Sample Treatment from empirical

- Let $\mathbf{P}_n$ be the probability distribution under which $(W, A) \sim P_n$ are sampled from empirical distribution, and Y, given W, A, are sampled from some estimate $q_{Y,n}(Y \mid W, A)$ of the true conditional distribution $q_{Y,0}$.
- One can evaluate the bias and variance and coverage of the estimation procedure based on repeated sampling of n i.i.d. observations from $\mathbf{P}_n$, **all w.r.t. truth** $\Psi(\mathbf{P}_n) = P_n E_{q_{Y,n}}(Y \mid A = 1, W)$.
- if $q_{Y,n}$ is a good estimator of $q_{Y,0}$ this could also be viewed as a **model based bootstrap** to construct confidence intervals in the actual data analysis.
- If we are in an outcome blind situation, $q_{Y,n}$ might be fitted on an external similar (qualitatively) data source or just set by the user.
- One might use such an **outcome blind simulation** to compare candidate estimators and pre-specify an estimation procedure for regulatory submission.

Bootstrap Approach 2: Generate Treatment from g_n

- Let $\tilde{P}_n$ be the probability distribution under which $W \sim P_n$, A, given W, has distribution g_n, and Y, given W, A is sampled from some estimate $q_{Y,n}(Y \mid W, A)$ of the true conditional distribution $q_{Y,0}$.
- As above, this could be used as a model based bootstrap for inference or as an outcome blind simulation study for comparing estimators or deciding on a pre-specified estimator.

Both model-based bootstrap methods are valid for inference if centered at estimator applied to true data distribution

- Let $P_n^{\#}$ be the empirical measure of a bootstrap sample from either "sample treatment distribution" $\mathbf{P}_n$ or "generate treatment distribution" $\tilde{P}_n$.
- For the sample-treatment bootstrap methods we have that

$$n^{1/2}(\hat{\Psi}(P_n^{\#}) - \hat{\Psi}(\mathbf{P}_n)) \Rightarrow_d N(0, \sigma^2)$$

  with the same normal limit distribution as $\hat{\Psi}(P_n)$, assuming the asymptotic normality

$$n^{1/2}(\hat{\Psi}(P_n) - \Psi(P_0)) \Rightarrow_d N(0, \sigma^2).$$

- The analogue applies to the generate-treatment bootstrap:

$$n^{1/2}(\hat{\Psi}(P_n^{\#}) - \hat{\Psi}(\tilde{P}_n)) \Rightarrow_d N(0, \sigma^2).$$

- Therefore, one can construct valid confidence intervals based on the lower and upper quantiles of these bootstrap distributions.

The "sample treatment" $\mathbf{P}_n$-bootstrap fails for simulations when centering estimator at "truth" $\Psi(\mathbf{P}_n)$

- Consider IPTW estimator. We have

$$\begin{aligned}\hat{\Psi}_{IPTW}(P_n^{\#}) - \Psi(\mathbf{P}_n) &= \hat{\Psi}_{IPTW}(P_n^{\#}) - \hat{\Psi}_{IPTW}(\mathbf{P}_n) \\ &\quad + \hat{\Psi}_{IPTW}(\mathbf{P}_n) - \Psi(\mathbf{P}_n) \\ &\sim N(0, \sigma^2) + \hat{\Psi}_{IPTW}(\mathbf{P}_n) - \Psi(\mathbf{P}_n).\end{aligned}$$

- Note that, contrary to $\hat{\Psi}_{IPTW}(\tilde{P}_n) - \Psi(\tilde{P}_n) = 0$, we dont have that $\hat{\Psi}_{IPTW}(\mathbf{P}_n) - \Psi(\mathbf{P}_n)$ equals zero.
- Specifically, the bias term is given by:

$$\begin{aligned}& b_n = \hat{\Psi}_{IPTW}(\mathbf{P}_n) - \Psi(\mathbf{P}_n) \\ & = P_n A/g_n(1 \mid W) E_{q_{Y,n}}(Y \mid A = 1, W) - P_n E_{q_{Y,n}}(Y \mid A = 1, W) \\ & = P_n E_{q_{Y,n}}(Y \mid A = 1, W)/g_n(1 \mid W)(A - g_n(1 \mid W)).\end{aligned}$$

Continuation ..

- This bias term can be analyzed and is asymptotically linear with a specified influence curve given by $E_{q_{Y,n}}(Y \mid A = 1, W)/g_0(1 \mid W)(A - g_0(1 \mid W))$ minus the influence curve of $\Phi(g_n) - \Phi(g_0) = P_0 E_{q_{Y,0}}(Y \mid A = 1, W)/g_0(1 \mid W)(g_n - g_0)(1 \mid W))$.
- Conditional on P_n, this represents a fixed bias of order $1/n^{1/2}$.
- Therefore, conditional on P_n, $n^{1/2}(\hat{\Psi}(P_n^{\#}) - \Psi(\mathbf{P}_n))$ behaves as a normal $N(b_n, \sigma^2)$ with a bias term b_n that does not go to zero.

ATE estimators of interest

- **Propensity score matching (Match)**. Uses the generalized full optimal matching algorithm with replacement (Hansen, 2004; Savje et al., 2021) to generate weights. The outcome model for $E(Y|A)$ is estimated using a weighted, unadjusted linear regression
- **Inverse probability of treatment weighting (IPTW)**. Weights stabilized by marginal treatment probability and bounded by $\sqrt{n}\ln(n)/5$ (Gruber et al 2022). The outcome model $E(Y|A)$ is estimated using a weighted, unadjusted linear regression
- **Doubly robust targeted maximum likelihood estimation (TMLE)**. The TMLE (van der laan and Rubin 2006) is fit using the correctly specified working models for the treatment propensity and outcome, bounding the treatment assignment probabilities by $5/(\sqrt{n}\ln(n))$.
- **Generalized linear model, correctly specified (glmCM)**. Outcome model $E(Y|A,W)$ is fit using correctly specified regression model.
- **Generalized linear model, adjusted for propensity score (glmPS)**. Outcome model is fit regressing Y on A and the PS fit using a correctly specified model $E(A|W)$.

Synthetic data simulations

General set-up

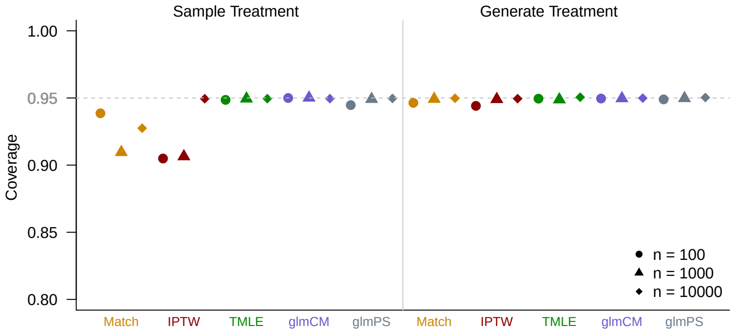
- Varied cohort size: $n = 100, 1000, 10000$
- Simple logistic binary treatment model, roughly 45% probability
  - Also considered a 1-1 randomized treatment for a few scenarios
- Simple generalized linear outcome models: continuous and binary
  - For binary outcome considered common (30%) and rare (5%) outcomes
- Compared performance of estimation methods for ATE
  - For binary outcome, also considered the relative risk (RR) and the conditional log OR (clogOR) from a marginal structural model
- 100,000 Monte Carlo simulation iterations
- Consider the mean bias, empirical SE, RMSE, and bias:SE ratio

Simulation: Estimate ATE for continuous outcome

$\psi_0^{ATE} = 2$

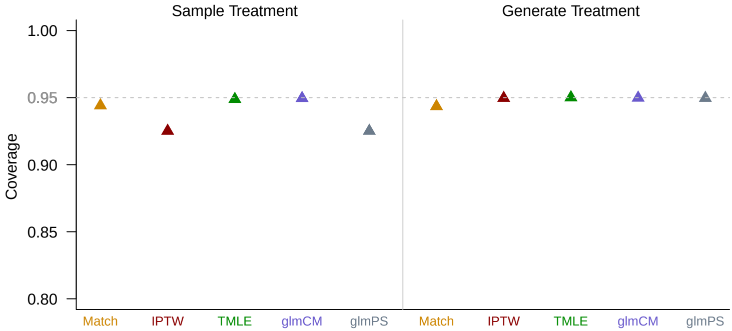
	Sample Treatment				Generate Treatment			
	% Bias	SE	RMSE	Bias:SE	% Bias	SE	RMSE	Bias:SE
$n = 100$								
Unadj	159.29	1.337	3.455	2.382	159.60	1.344	3.463	2.376
Match	−10.78	0.818	0.846	0.264	3.80	0.843	0.846	0.090
IPTW	−19.65	0.612	0.727	0.642	2.08	0.478	0.479	0.087
TMLE	−0.15	0.236	0.236	0.013	−0.15	0.234	0.234	0.013
glmCM	−0.01	0.224	0.224	0.001	0.01	0.223	0.223	0.001
glmPS	−2.41	0.248	0.252	0.195	0.10	0.236	0.236	0.009
$n = 1000$								
Unadj	19.53	0.470	0.611	0.831	19.33	0.469	0.608	0.824
Match	−18.01	0.548	0.655	0.658	0.85	0.419	0.419	0.041
IPTW	−7.28	0.235	0.276	0.620	0.18	0.221	0.221	0.016
TMLE	−0.13	0.076	0.076	0.034	−0.06	0.076	0.076	0.016
glmCM	0.00	0.071	0.071	0.001	−0.01	0.071	0.071	0.002
glmPS	−0.08	0.072	0.072	0.021	0.00	0.071	0.071	0.001
$n = 10000$								
Unadj	43.62	0.141	0.884	6.168	43.57	0.142	0.883	6.138
Match	3.43	0.154	0.169	0.445	0.00	0.117	0.117	0.001
IPTW	0.09	0.056	0.056	0.033	0.01	0.058	0.058	0.003
TMLE	0.00	0.023	0.023	0.003	0.00	0.024	0.024	0.004
glmCM	0.00	0.022	0.022	0.002	0.00	0.022	0.022	0.001
glmPS	−0.01	0.022	0.022	0.009	0.00	0.022	0.022	0.001

Problematic coverage: Continuous outcome



Problematic coverage: Continuous outcome, Randomized treatment

$n = 1,000$, 1:1 randomization



Simulation: Estimate ATE for binary outcome

$\psi_0^{ATE} = 0.2199, 0.2171, 0.2182$, when $n = 100$, 1000, 10,000, respectively

	Sample Treatment				**Generate Treatment**			
	% Bias	SE	RMSE	Bias:SE	% Bias	SE	RMSE	Bias:SE
$n = 100$								
Unadj	29.25	0.091	0.111	0.708	29.71	0.091	0.112	0.720
Match	1.15	0.129	0.129	0.020	1.11	0.119	0.119	0.020
IPTW	−2.23	0.110	0.110	0.044	0.51	0.106	0.106	0.011
TMLE	0.11	0.106	0.106	0.002	0.16	0.106	0.106	0.003
glmCM	0.10	0.101	0.101	0.002	0.18	0.101	0.101	0.004
glmPS	−0.34	0.101	0.101	0.007	−0.02	0.101	0.101	0.000
$n = 1000$								
Unadj	32.81	0.028	0.077	2.538	33.05	0.028	0.077	2.556
Match	0.13	0.043	0.043	0.006	0.27	0.039	0.039	0.015
IPTW	−0.66	0.034	0.034	0.042	0.07	0.034	0.034	0.005
TMLE	−0.06	0.034	0.034	0.004	0.00	0.034	0.034	0.000
glmCM	−0.05	0.032	0.032	0.004	0.00	0.032	0.032	0.000
glmPS	−0.15	0.032	0.032	0.010	−0.05	0.032	0.032	0.003
$n = 10000$								
Unadj	32.18	0.009	0.071	7.909	32.22	0.009	0.071	7.862
Match	0.29	0.013	0.013	0.048	0.03	0.012	0.012	0.006
IPTW	0.34	0.010	0.010	0.071	0.02	0.011	0.011	0.005
TMLE	0.02	0.010	0.010	0.003	0.02	0.011	0.011	0.004
glmCM	0.01	0.010	0.010	0.001	0.01	0.010	0.010	0.003
glmPS	−0.03	0.010	0.010	0.006	0.00	0.010	0.010	0.001

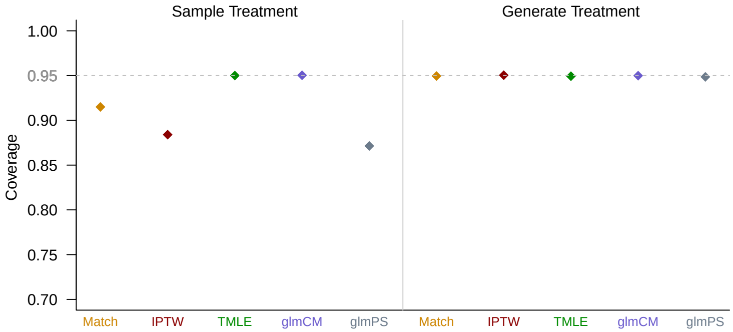
SE: Standard Error; RMSE: root mean squared error

Simulation: Estimate ATE for rare binary outcome

$\psi_0^{ATE} = -0.0247$, $n = 10,000$, 5% outcome rate

	Sample Treatment				Generate Treatment			
	% Bias	SE	RMSE	Bias:SE	% Bias	SE	RMSE	Bias:SE
Unadj	41.698	0.003	0.011	3.369	34.918	0.003	0.009	2.793
Match	8.493	0.004	0.004	0.555	−0.103	0.003	0.003	0.007
IPTW	9.362	0.003	0.004	0.747	−0.052	0.003	0.003	0.004
TMLE	−0.001	0.003	0.003	0.000	−0.052	0.003	0.003	0.005
glmCM	−0.044	0.002	0.002	0.004	−0.033	0.002	0.002	0.003
glmPS	9.939	0.003	0.004	0.813	1.343	0.003	0.003	0.108

Problematic Coverage for ATE: Rare binary outcome



Simulation: Estimate logcOR when MSM is not equivalent to underlying outcome model

True outcome model (14% probability):
$\text{logit}(P(Y=1|A,\mathbf{W}) = \beta_0 + \beta_1 A + \beta_2 W_1 + \beta_3 W_2 + \beta_4 W_3 + \beta_5 W_4 + \beta_6 W_5$

MSM model: incorrect logistic regression that omitted (W_4, W_5), logcOR $= 1.084$

	Sample Treatment				**Generate Treatment**			
	% Bias	SE	RMSE	Bias:SE	% Bias	SE	RMSE	Bias:SE
$n = 100$	60.424	2.829	2.904	0.232	50.114	2.712	2.766	0.200
$n = 1000$	4.189	0.228	0.232	0.199	1.323	0.229	0.229	0.063
$n = 10000$	0.780	0.071	0.072	0.118	0.104	0.071	0.071	0.016

Real data example

- Kaiser Permanente Washington (KPWA) is an integrated health care system in Pacific Northwest that provides care and health insurance to over 700,000 members
- 112,770 KPWA adults aged 13+ years, initiating antidepressant medication or psychotherapy from January 1, 2008 to December 31 2018 (n=112,770)
  - No antidepressant fills or psychotherapy in the prior year
- Plasmode data set: 50,337 individuals with complete data on the Patient Health Questionnaire (PHQ-9)
- Outcome: Composite outcome of self-harm (fatal or non-fatal) or psychiatric hospitalization within 5 years following treatment initiation n=5193, (10.3%)

Plasmode simulation

Confounders bootstrapped sampled from KPWA Cohort

- N=10,000

Data generating Models for treatment and outcome

- Binary treatment data generating model - logistic
  - Antidepressant medication or psychotherapy
- Binary outcome data generating model - logistic
  - Self-harm/Psychiatric hospitalization within 5 years of treatment initiation

Model parameters estimated from KPWA Cohort

- Treatment and outcome model fit to 50,337 with complete data
- For each type of generating model use KPWA cohort to estimate logistic regression model with interactions
- For simplicity, analysis model matched the data generating model

KPWA-based logistic models: real data and data generating models for 15% and 5% outcomes

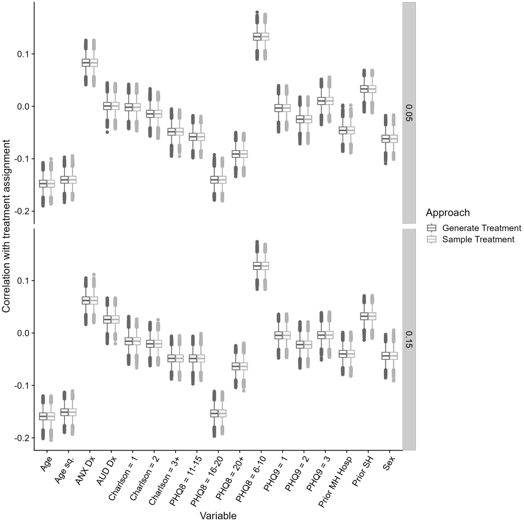
Variable	Receipt of PT	5-year SH/HOSP	15% outcome	5% outcome
Intercept	2.361	-2.063	-1.320	-2.350
Psychotherapy	NA	-0.206	-1.000	-3.100
Female sex	-0.238	0.360	0.360	0.360
Age at initiation	-0.030	-0.060	-0.060	-0.060
Charlson 1	-0.041	0.176	0.176	0.176
Charlson 2	0.084	0.953	0.953	0.953
Charlson 3+	0.907	1.988	1.988	1.988
Alcohol use disorder	0.242	0.842	0.842	0.842
Anxiety disorder	0.454	0.096	0.096	0.096
Prior self-harm	0.145	1.960	1.960	1.960
Prior hospitalization with MH diagnosis	-0.320	0.914	0.914	0.914
PHQ8: 6–10	-0.878	-0.026	-0.026	-0.026
PHQ8: 11–15	-1.674	0.209	0.209	0.209
PHQ8: 16–20	-2.074	0.338	0.338	0.338
PHQ8: 21–24	-2.126	0.349	0.349	0.349
PHQ9: 1	0.139	0.222	0.222	0.222
PHQ9: 2	0.118	0.296	0.296	0.296
PHQ9: 3	0.450	0.548	0.548	0.548
Age at initiation squared	0.000	0.001	0.001	0.001
Charlson score 1 & anxiety disorder	-0.090	-0.180	-0.180	-0.180
Charlson score 2 & anxiety disorder	0.298	0.146	0.146	0.146
Charlson score 3+ & anxiety disorder	0.033	0.260	0.260	0.260
Age at initiation & female sex	0.000	-0.007	-0.007	-0.007
Female sex & prior self-harm	0.155	-0.014	-0.014	-0.014
Age at initiation & prior self-harm	-0.003	-0.020	-0.020	-0.020
Charlson score 1 & age at initiation	0.002	0.002	0.002	0.002
Charlson score 2 & age at initiation	-0.001	-0.007	-0.007	-0.007
Charlson score 3+ & age at initiation	-0.013	-0.019	-0.019	-0.019
PHQ item 9 score 1 & female sex	0.085	-0.042	-0.042	-0.042
PHQ item 9 score 2 & female sex	0.051	-0.064	-0.064	-0.064
PHQ item 9 score 3 & female sex	0.026	0.059	0.059	0.059
PHQ item 9 score 1 & prior self-harm	0.497	-0.218	-0.218	-0.218
PHQ item 9 score 2 & prior self-harm	0.889	-0.494	-0.494	-0.494
PHQ item 9 score 3 & prior self-harm	0.330	-0.534	-0.534	-0.534

KPWA Simulation Results for the 5% outcome

$\psi_0^{ATE} = -0.079$, $\psi_0^{RR} = 0.062$, $n = 10,000$

		Sample Treatment					Generate Treatment				
Estimand	Estimator	% Bias	SE	RMSE	bias:SE	CP	% Bias	SE	RMSE	bias:SE	CP
ATE	Unadj	10.964	0.004	0.010	2.130	43.4	11.191	0.004	0.010	2.169	41.8
	Match	0.403	0.005	0.005	0.064	94.9	-0.245	0.005	0.005	0.042	95.0
	IPTW	1.189	0.005	0.005	0.195	95.3	-0.219	0.004	0.004	0.042	95.1
	TMLE	0.571	0.005	0.005	0.096	95.1	0.012	0.004	0.004	0.002	95.1
	glmCM	-0.175	0.004	0.004	0.034	95.3	-0.182	0.004	0.004	0.036	95.1
	glmPS	-2.553	0.004	0.004	0.519	91.9	-2.874	0.004	0.004	0.586	90.9
RR	Unadj	-20.563	0.011	0.017	1.166	77.9	-20.875	0.011	0.017	1.188	77.3
	Match	-3.705	0.019	0.019	0.123	95.6	-1.548	0.018	0.018	0.054	95.5
	IPTW	0.328	0.016	0.016	0.013	95.2	0.340	0.016	0.016	0.014	95.2
	TMLE	-0.555	0.016	0.016	0.022	95.2	0.062	0.016	0.016	0.002	95.1
	glmCM	0.362	0.014	0.014	0.016	95.1	0.326	0.014	0.014	0.014	95.1
	glmPS	4.675	0.014	0.015	0.201	94.6	5.558	0.015	0.015	0.237	94.3

Correlation between treatment and covariates in KPWA simulation



Conclusions

- One could carry out a model based bootstrap for inference with both Sample Treatment ($\mathbf{P}_n$) and Generate Treatment ($\tilde{P}_n$) approaches.
- However, evaluation of the sampling distribution of $n^{1/2}(\hat{\Psi}(P_n^{\#}) - \Psi(\mathbf{P}_n))$ is biased w.r.t. $n^{1/2}(\hat{\Psi}(P_n) - \Psi(P_0))$ even if $q_{Y,n}$ is consistent for $q_{Y,0}$.
  - Bias is negligile for a pure outcome regression based estimator.
  - Bias is non-negligible (as large as $n^{-1/2}$) for an IPTW or double robust estimator that does not want to fully rely on correct estimation of the outcome regression.
- If one uses machine learning to estimate g_0, then the $\mathbf{P}_n$-bootstrap could be inconsistent, while the $\tilde{P}_n$-bootstrap will still be consistent.
- The Generate Treatment and Sample Treatment algorithm can similarly approximate the desired data features
- We recommend the Generate Treatment $\tilde{P}_n$-bootstrap for simulation studies.