From SATE to PATT: Combining Experimental with Observational Studies to Estimate Population Treatment Effects

Erin Hartman Richard Grieve Roland Ramsahai

Jasjeet S. Sekhon

May 17, 2013

The Problem

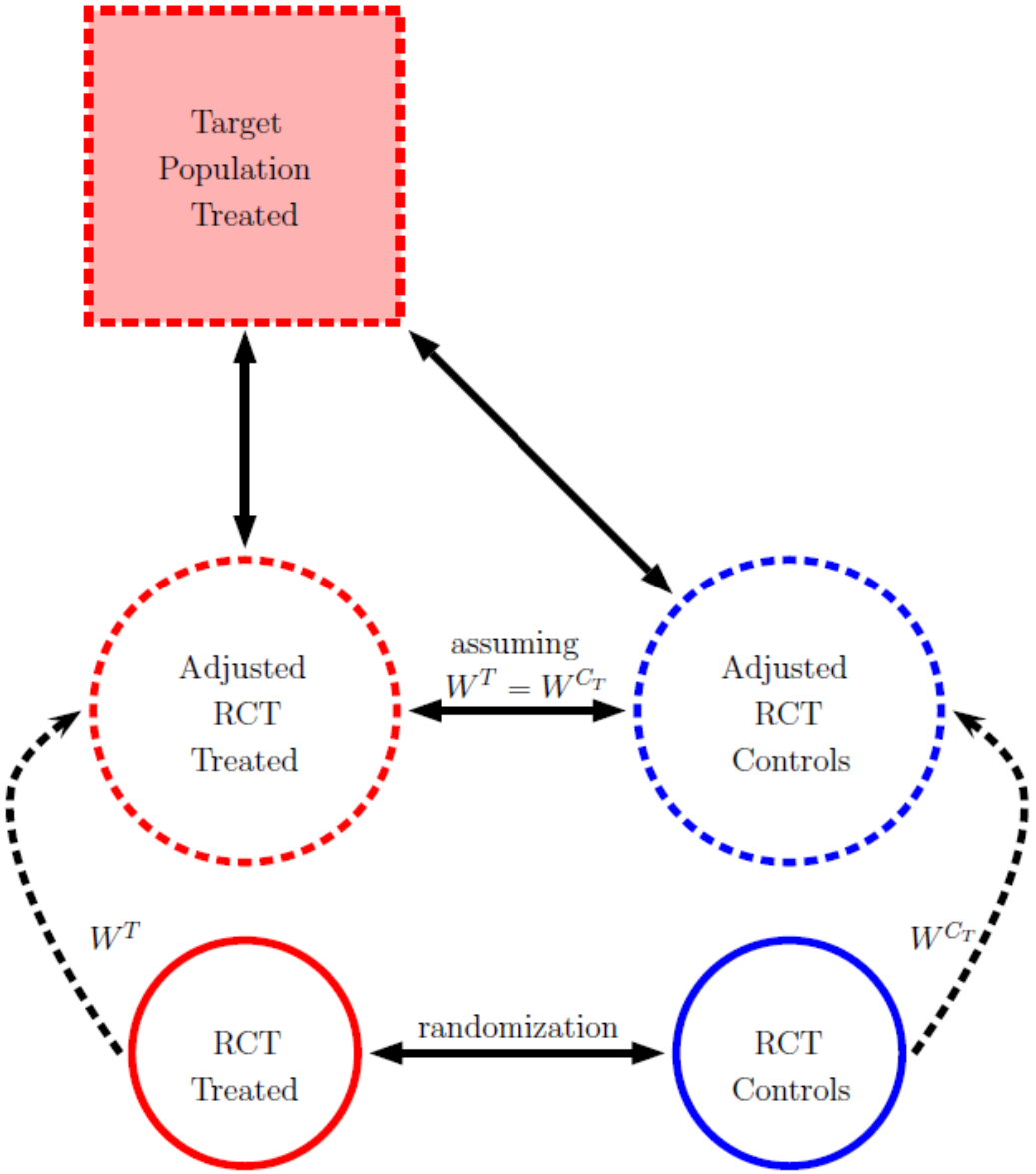
- Randomized Controlled Trials (RCTs) allow one to estimate treatment effects with few assumptions
- RCTs are usually not conducted on the population of interest; Non-Random Studies (NRSs) sometimes are
- Problem: how to combine information from both RCTs and NRSs to estimate treatment effects in the population?
  - RCTs raise issues of Randomization Bias: poor external validity
  - NRSs raise issues of selection bias, or non random assignment to treatment: poor internal validity

Example

- A growing interest in the cost of health care
- For costs, there may be a bigger gap between population and sample treatment effects than for clinical outcomes
- e.g., compare the cost of UCLA versus Mayo Clinic; Dartmouth Atlas of Health Care
- We examine a simple case: conduct a cost effectiveness analysis (CEA) for one medical procedure

Estimands

- RCTs allow us to identify the Sample Average Treatment Effect (SATE), which is asymptotically equivalent to the Sample Average Treatment Effect on the Treated (SATT)
- We are often interested in the treatment effect for those who would receive treatment in practice, or the Population Average Treatment Effect on the Treated (PATT)


Target
Population
Treated
Adjusted
RCT
Treated
assuming
$W^T = W^{C_T}$
Adjusted
RCT
Controls
W^T
W^{C_T}
RCT
Treated
randomization
RCT
Controls

Our Method

- Introduce new designs for combining RCTs and NRSs
  - Using stratification or matching
    - SATE $\rightarrow$ SATT
  - Using Maximum Entropy Weighting to maximize the external validity
    - SATT $\rightarrow$ PATT
- One design, similar to difference-in-difference, does not require a selection on observables assumption
- Importantly: we provide placebo tests to validate the identifying assumptions

Pulmonary Artery Catheterization (PAC)

- PAC is an invasive cardiac monitoring device for critical ill patients (ICU)—e.g., myocardial infarction (ischaemic heart disease)
- Widely used for the past 30 years: spend $2 billion in U.S. per year
- RCTs find no effect; seven NRS find that PAC increases mortality (e.g., Connors et al. JAMA 1996)

Pulmonary Artery Catheterization

- RCT: a publicly funded, pragmatic experiment done in 65 UK ICUs in 2000-2004.
  - 1014 subjects, 506 who received PAC
  - No difference in hospital mortality ($p = 0.39$)
- NRS: all ICU admissions to 57 UK ICUs in 2003–2004
  - 1052 cases with PAC and 32,499 controls
  - One observational study was able to find no difference in hospital mortality ($p = 0.29$)
- However, the populations (and centers) between the two studies differ, and we are interested in identifying PATT

Results and Implications

- PAC helps patients undergoing elective surgery, but not the overall population
- We pass placebo tests for both hospital mortality and costs, as well as measures of cost-effectiveness
- NRSs could not recover experimental benchmarks for costs not because of sample selection bias, but because of confounding of treatment selection bias
- Has implications for cost-effectiveness analysis, which often relies on observational studies to predict long term outcomes and estimate costs in routine clinical practice

The Method

Let θ_s be a stratified treatment effect from the randomized trial, θ_{ws} be the treatment effect of reweighted strata, and θ be the true population treatment effect.

$$\theta_s \rightarrow \theta_{ws} \rightarrow \theta$$

- Estimate θ_s using Genetic Matching
- Reweight the strata using Maximum Entropy weighting to estimate θ_{ws}
- Run a placebo test to validate the identifying assumptions and provide evidence for how close θ_{ws} is to θ

Some Definitions

- Let W denote a set of conditioning covariates, with the distribution of the population treated observation
- Let $Y_{s,t}$ denote the potential outcomes for a given subject in sample s and treatment t
- Let $T \in (0, 1)$ be an indicator for whether or not subject i was in the treatment or control group
- Let $S \in (0, 1)$ be an indicator for whether or not subject i was in the sample population

Testing the Assumptions

A.1 Treatment is consistent across studies:
$Y_{i01} = Y_{i11}$ and $Y_{i00} = Y_{i10}$

A.2 Strong Ignorability of Sample Assignment for Treated:

$$(Y_{i01}, Y_{i11}) \perp\!\!\!\perp S_i | (W_i, T_i = 1) \; 0 < Pr(S_i = 1 | W_i, T_i = 1) < 1$$

A.3 Strong Ignorability of Sample Assignment for Controls:

$$(Y_{i00}, Y_{i10}) \perp\!\!\!\perp S_i | (W_i, T_i = 1) \; 0 < Pr(S_i = 1 | W_i, T_i = 1) < 1$$

A.4 SUTVA: The potential outcomes for a given individual do not depend on the treatment assignment of other individuals

Placebo Test

- Natural placebo test arises:
  - The difference between the mean outcome of the NRS treated and mean outcome of the reweighted RCT treated should be zero
  - If not, then at least one of the assumptions has failed
- There is a similar placebo test for controls, however, it does not provide as much information
  - Could fail due to lack of overlap, for example
  - Failure is not as informative about accuracy of the PATT estimate as the treated placebo

Placebo Test

- Natural placebo test arises:
  - The difference between the mean outcome of the NRS treated and mean outcome of the reweighted RCT treated should be zero
  - If not, then at least one of the assumptions has failed
- There is a similar placebo test for controls, however, it does not provide as much information
  - Could fail due to lack of overlap, for example
  - Failure is not as informative about accuracy of the PATT estimate as the treated placebo

Difference-in-Difference

An alternative design:

$$\tau_{PATT_{DID}} = \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i, S_i = 1, T_i = 1) - \mathbb{E}(Y_i|W_i, S_i = 1, T_i = 0)\}$$
$$-[\mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i, S_i = 1, T_i = 1)\} - E(Y|S_i = 0, T_i = 1)]$$

- The first difference is the adjusted experimental estimand and is intuitively a measure of the adjusted average effect
- The second difference is defined as the difference between the RCT treated and NRS treated
- Required that A3, A4, and part of A1 hold ($Y_{i00} = Y_{i10}$)

Statistical Methods

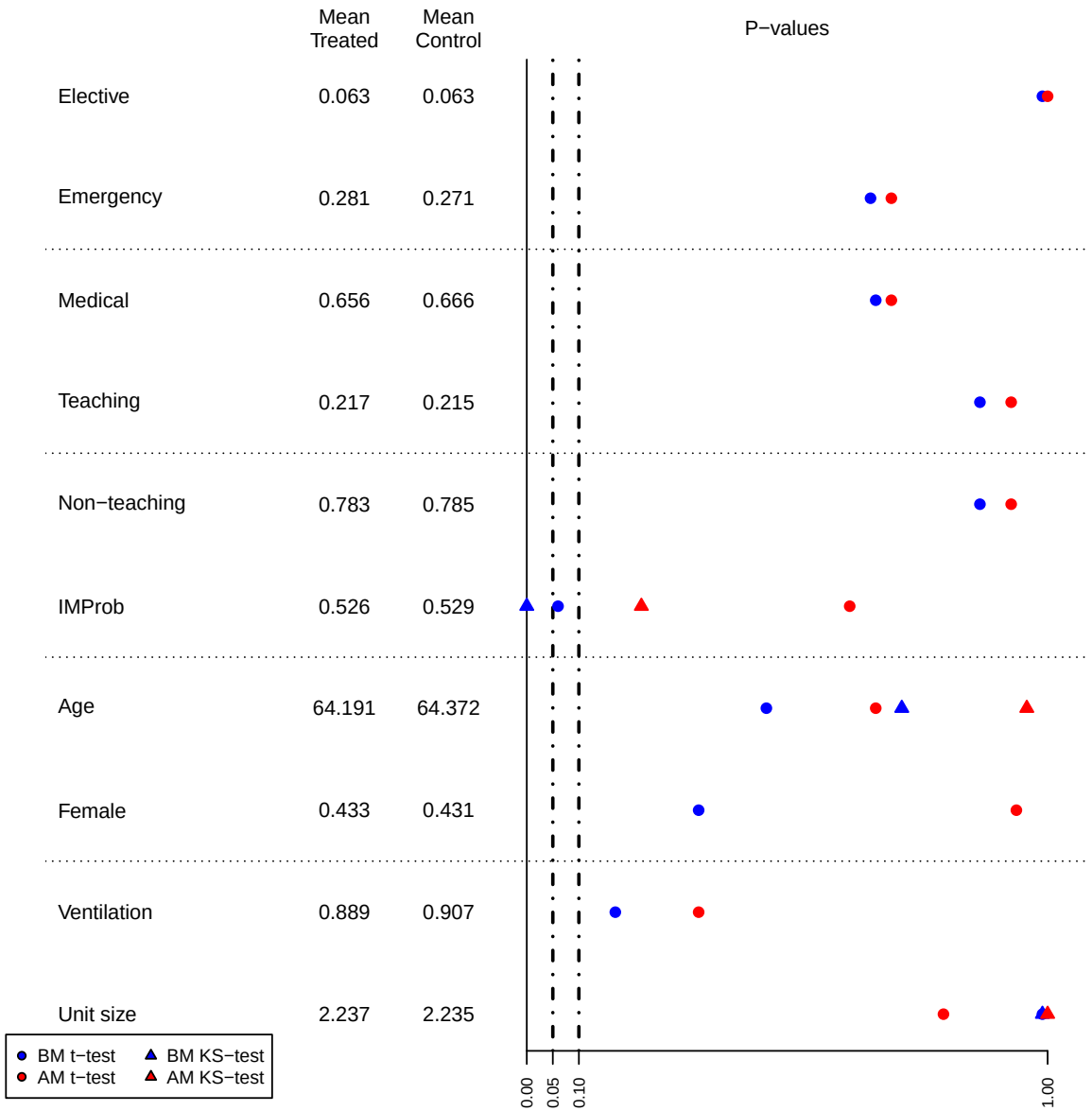
GenMatch:

- Matching method with automated balance optimization

Maximum Entropy weighting:

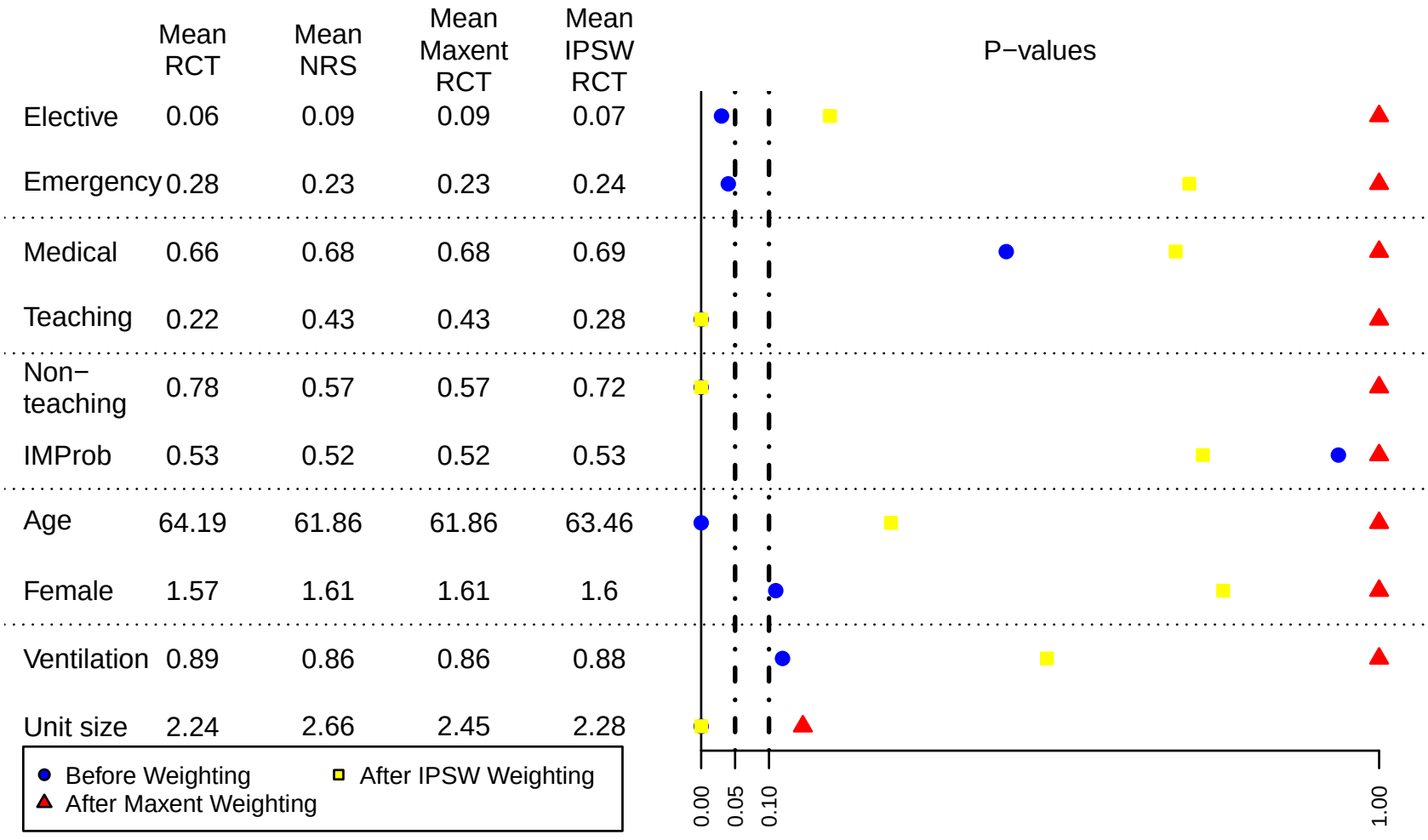
- Weighting method that assigns weights such that they simultaneously meet a set of consistency constraints while maximizing Shannon's measure of entropy
- Consistency constraints are based on moments of the population based on the NRS

Covariate Balance in RCT

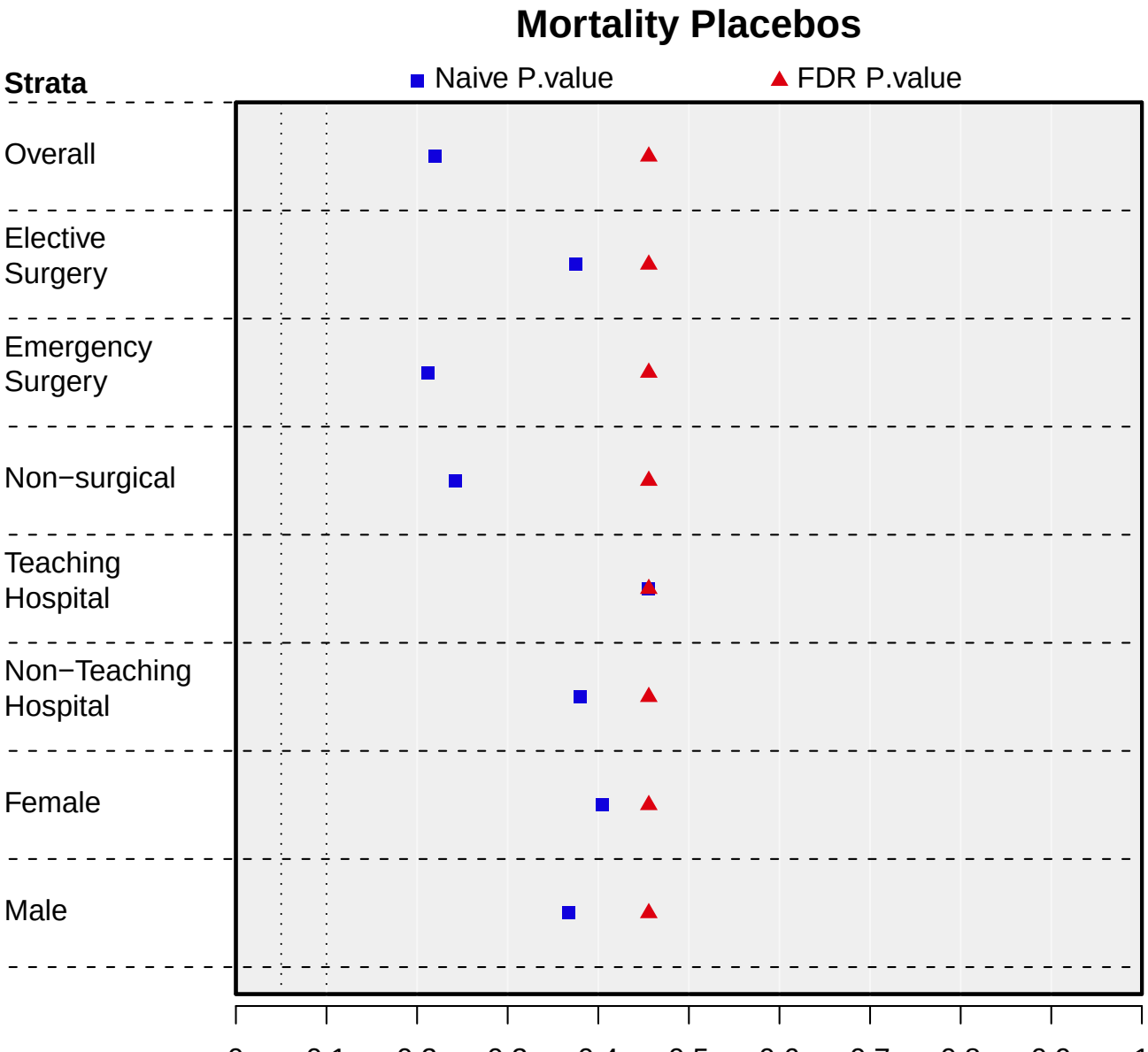


Balance Before and After Adjustment

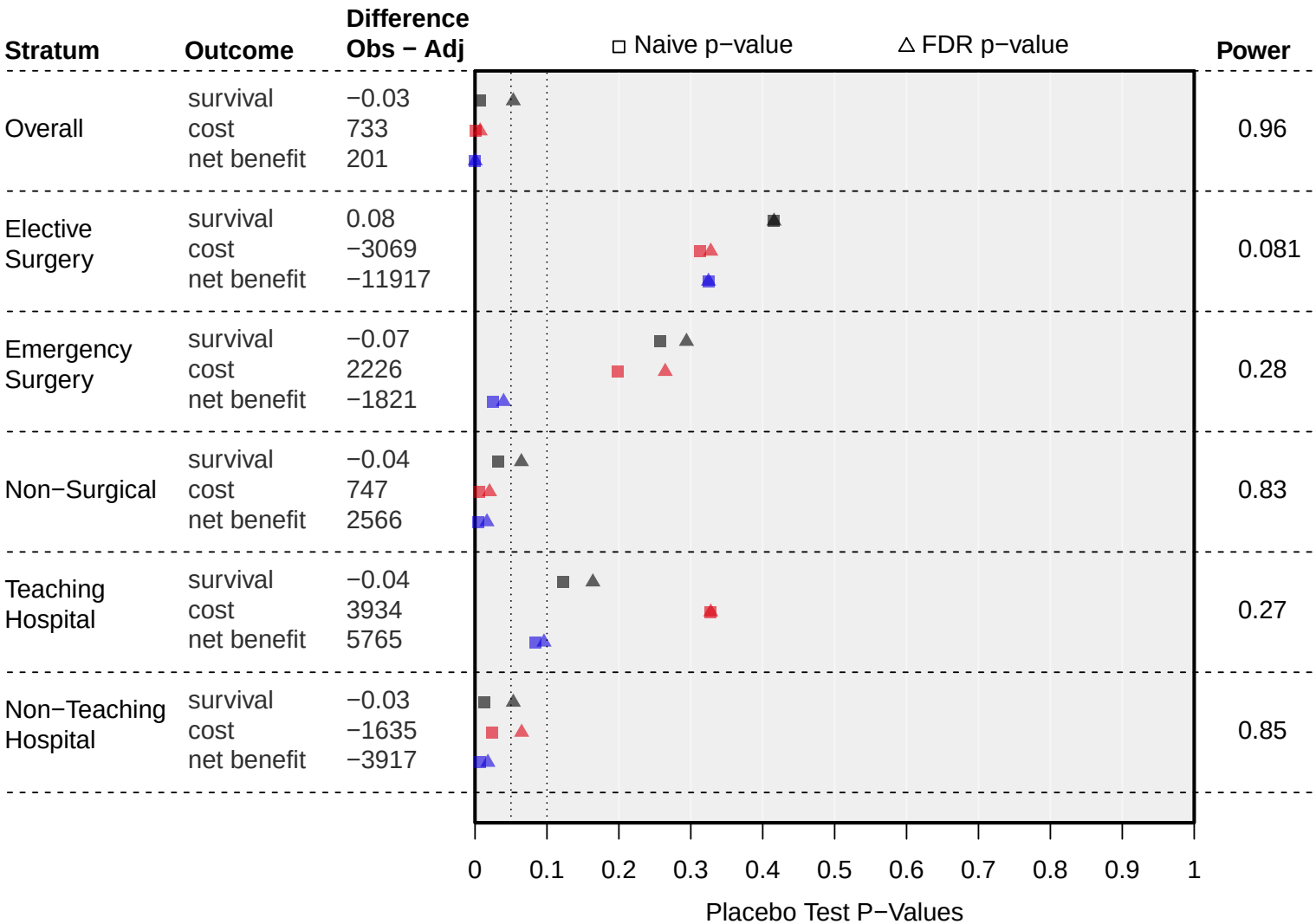
	Mean RCT	Mean NRS	Mean Maxent RCT	Mean IPSW RCT
Elective	0.06	0.09	0.09	0.07
Emergency	0.28	0.23	0.23	0.24
Medical	0.66	0.68	0.68	0.69
Teaching	0.22	0.43	0.43	0.28
Non−teaching	0.78	0.57	0.57	0.72
IMProb	0.53	0.52	0.52	0.53
Age	64.19	61.86	61.86	63.46
Female	1.57	1.61	1.61	1.6
Ventilation	0.89	0.86	0.86	0.88
Unit size	2.24	2.66	2.45	2.28



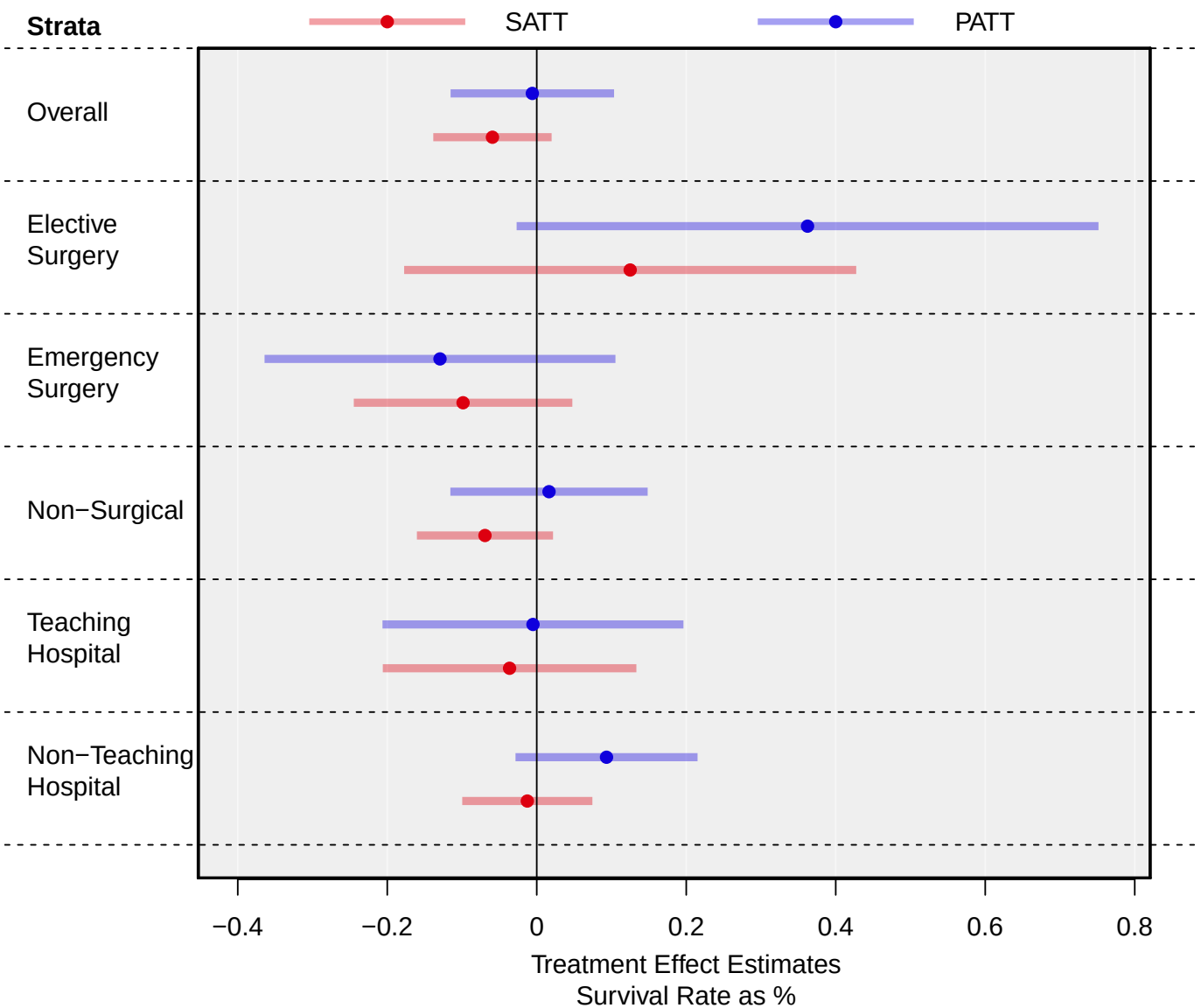
Placebo Tests



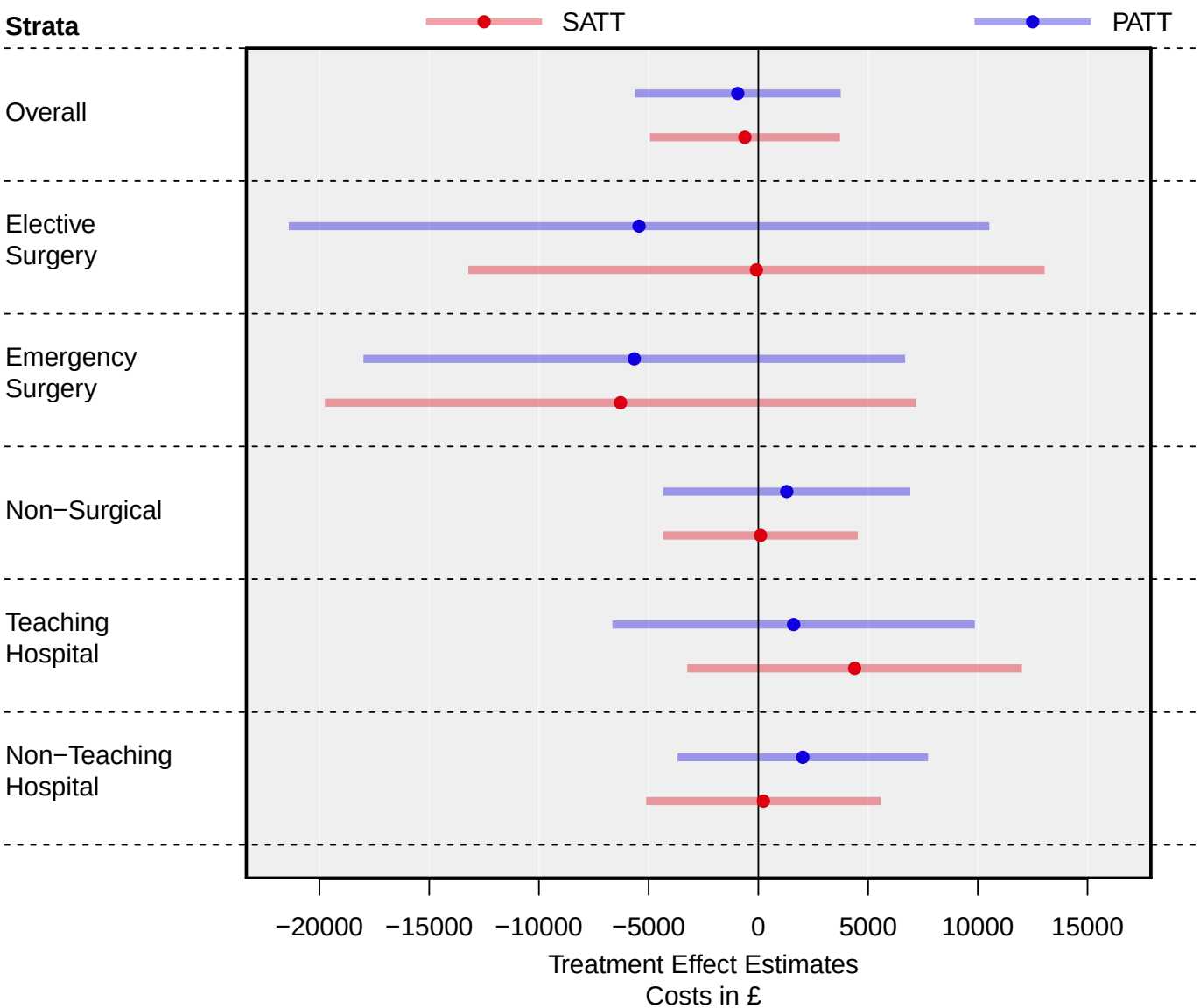
Placebo Tests



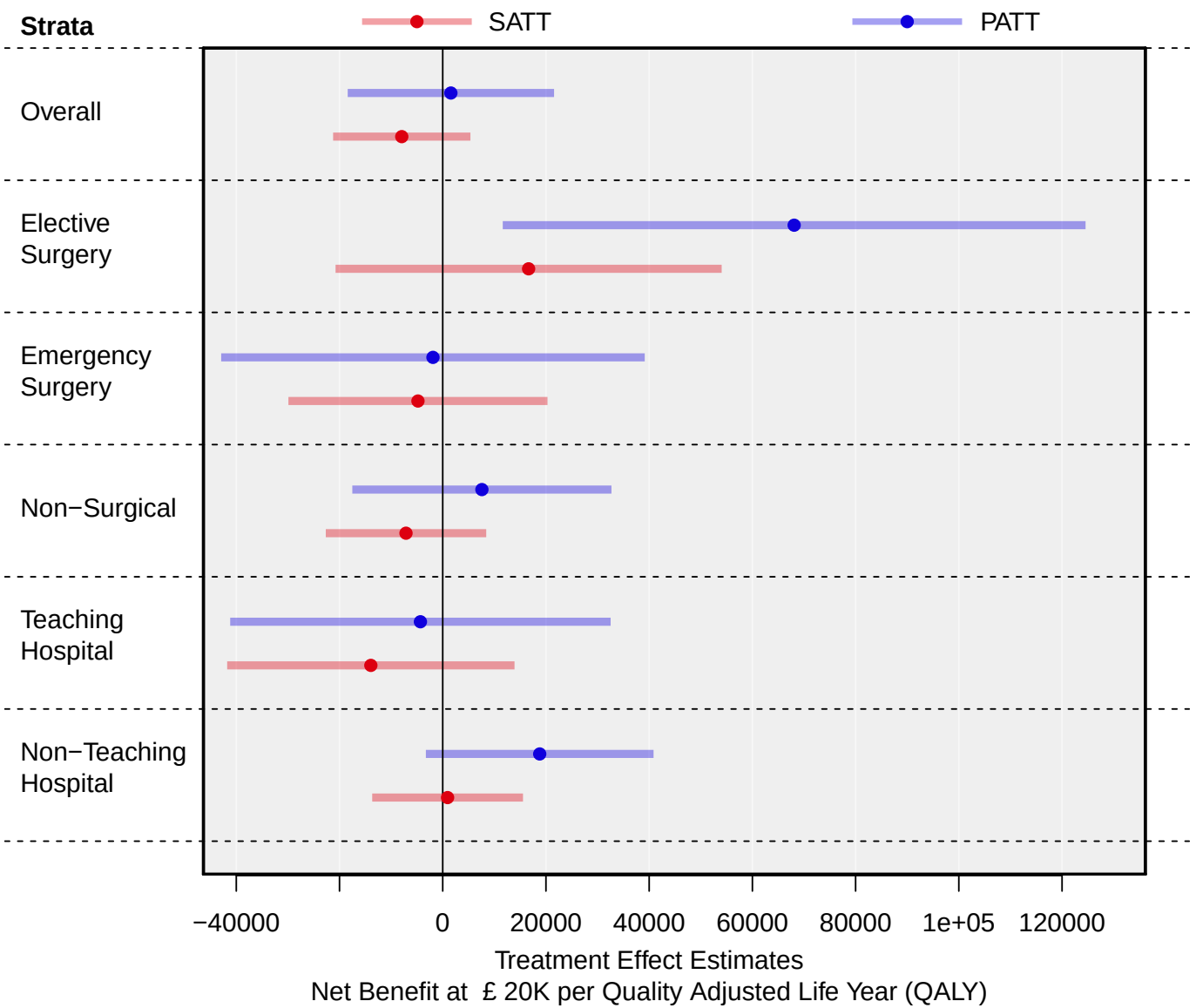
Population Treatment Effects on Hospital Survival Rates



Population Treatment Effects on Costs



Population Treatment Effects on Cost-Effectiveness



Conclusion

- Our method can be used to compare across different experiments
- The difference-in-difference probably of most general use
- Treated population estimand difficult to estimate without assumptions
- Only one observational method, Sekhon and Grieve, was able to recover the experimental benchmark for hospital mortality rates, and all observational methods were unable to do so for costs
  - This is due to treatment selection issues, not sample selection issues

Decomposition of Bias

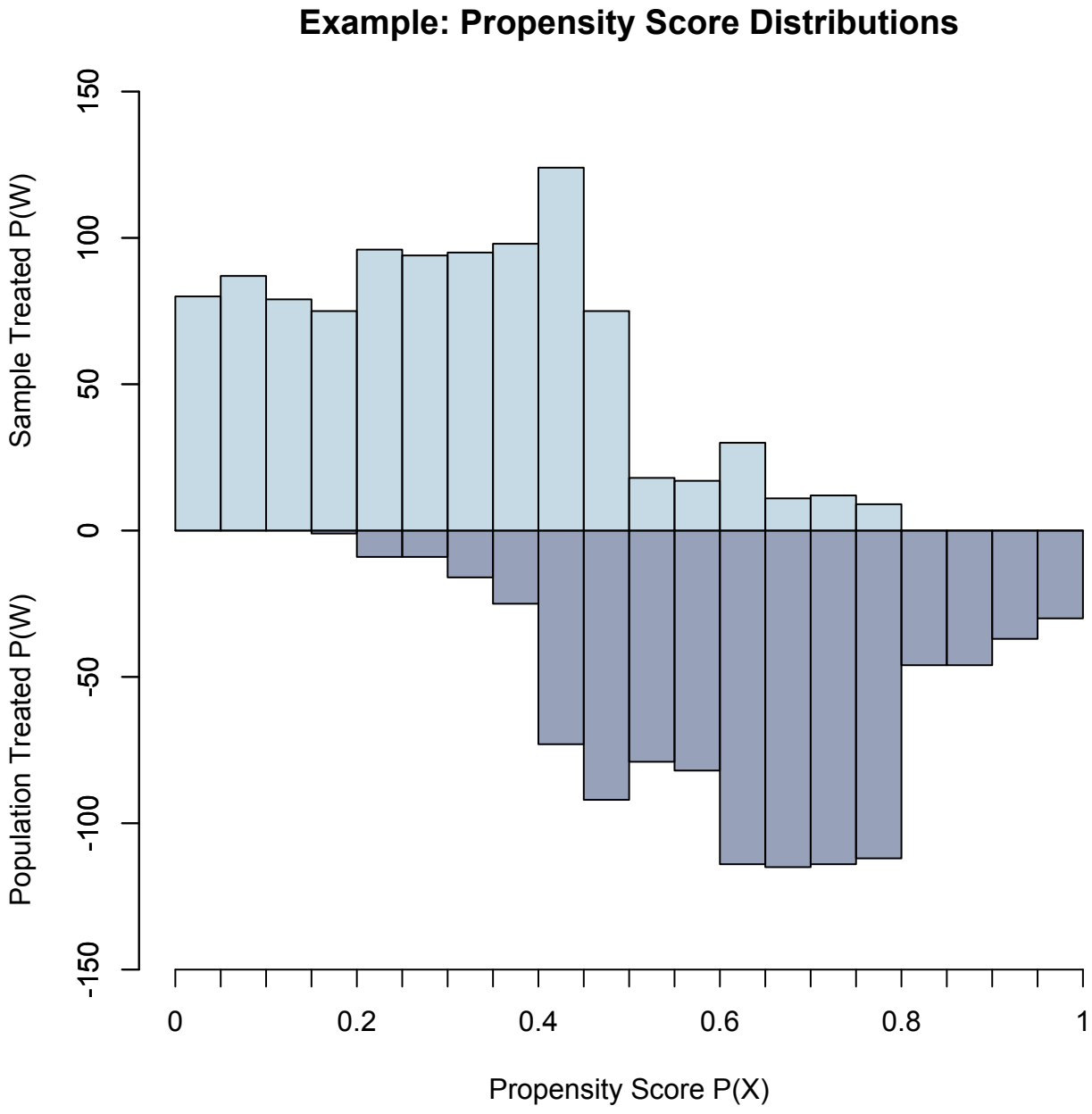
We can think of the bias in our estimate of the PATT as a combination of bias due to a possible lack of internal validity of the estimate of SATT and bias due to poor external validity of the estimate from the RCT.

$$B = B_I + B_E$$

$$
\begin{aligned}
B_E = \mathbb{E}_{S_{11} \setminus S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 1)\} \\
- \mathbb{E}_{S_{01} \setminus S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 0)\} \\
+ \mathbb{E}_{dF(S=1) - dF(S=0), S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 0)\} \\
+ \mathbb{E}_{dF(S=1), S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 1) - \mathbb{E}(Y_i | W, T = 1, S = 0)\} \\
\text{for } i \in (0, 1)
\end{aligned}
$$

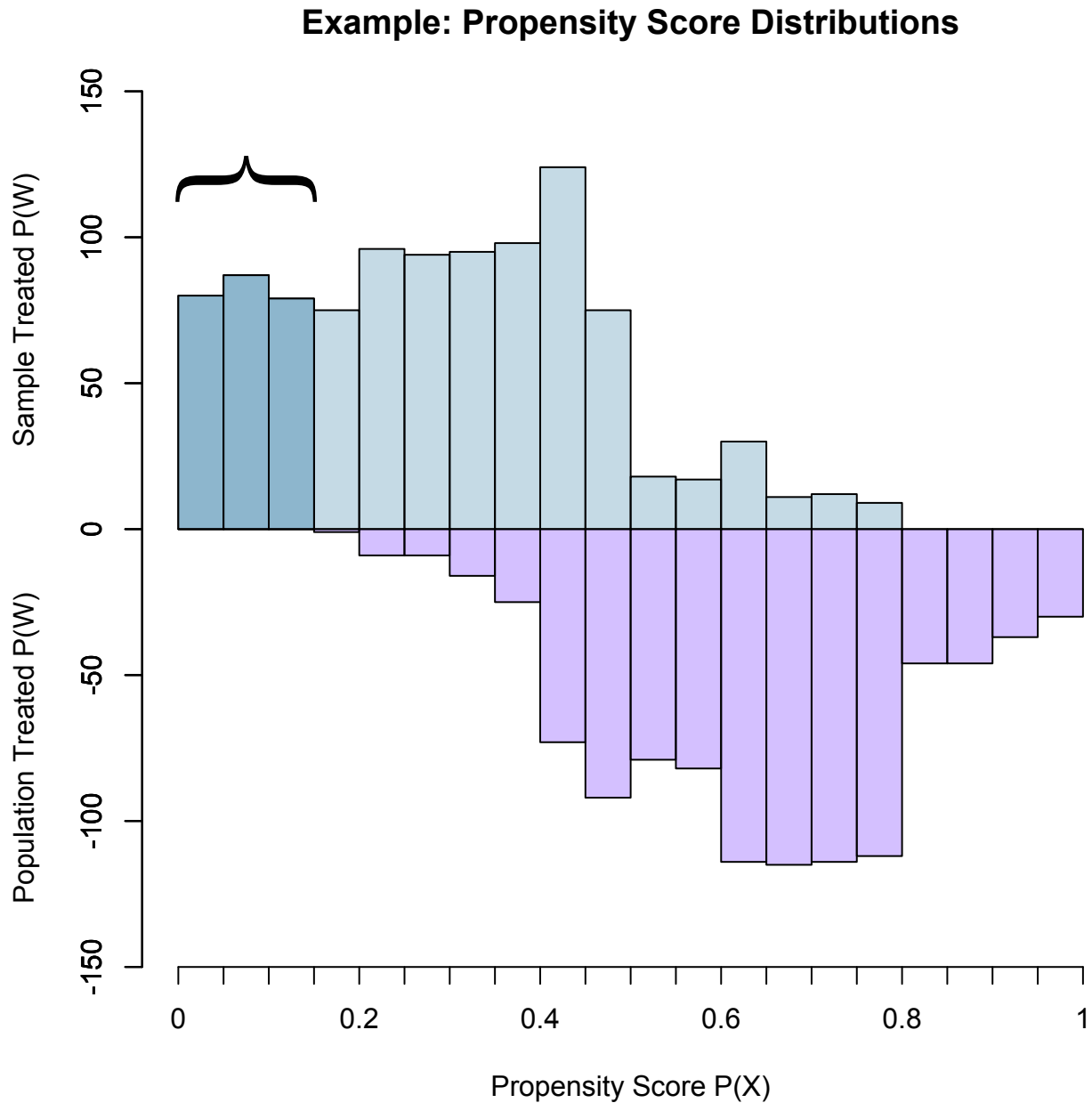
External validity problem:

- RCT is not a random sample of the population.
- Treatment can mean something different between the RCT and NRS.


Example: Propensity Score Distributions
Sample Treated P(W)
Population Treated P(W)
150
100
50
0
-50
-100
-150
0
0.2
0.4
0.6
0.8
1
Propensity Score P(X)

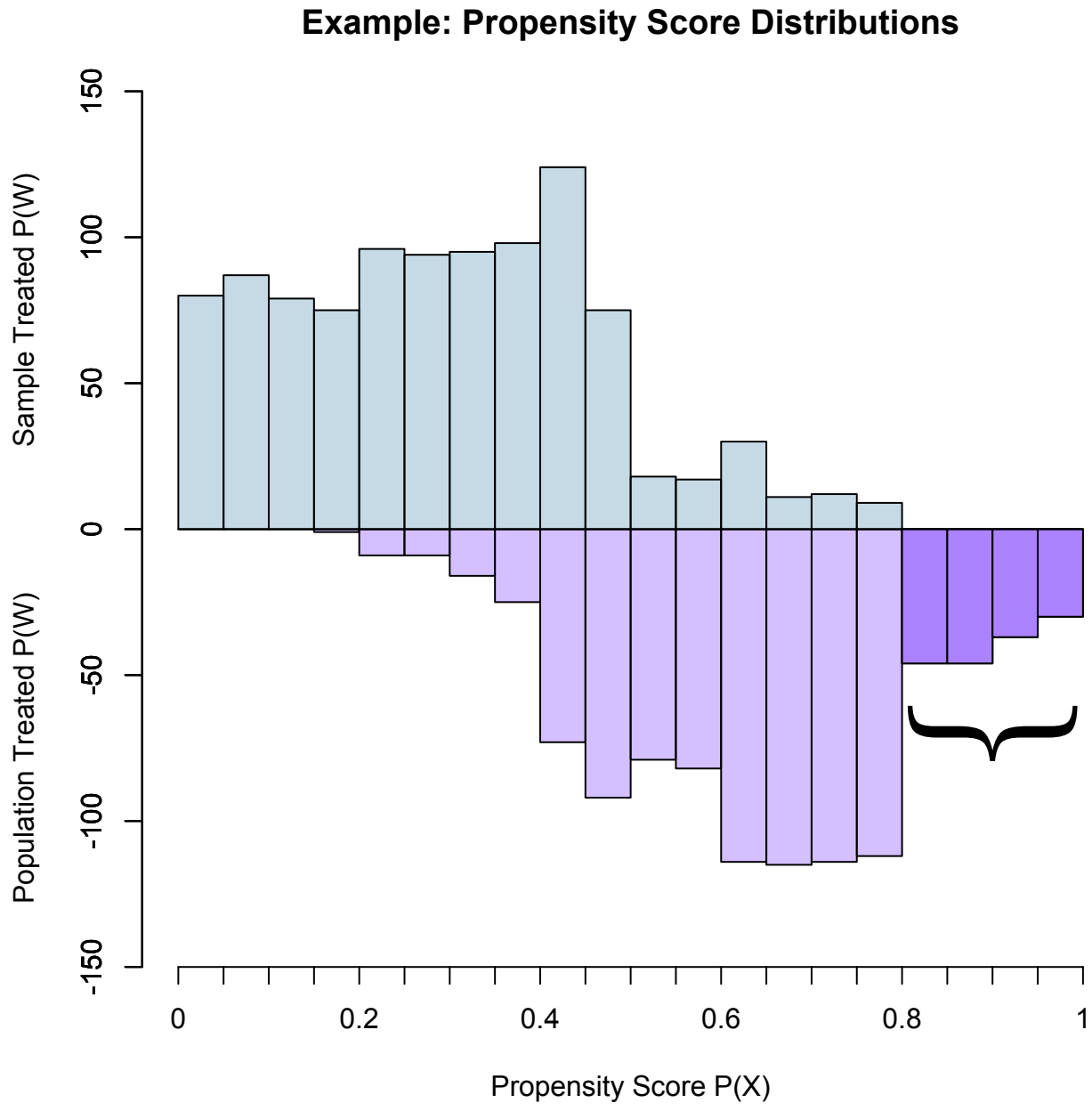
$$B_E = \mathbb{E}_{S_{11} \setminus S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 1)\}$$
$$- \mathbb{E}_{S_{01} \setminus S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 0)\}$$
$$+ \mathbb{E}_{dF(S=1) - dF(S=0), S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 0)\}$$
$$+ \mathbb{E}_{dF(S=1), S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 1) - \mathbb{E}(Y_i | W, T = 1, S = 0)\}$$
$$\text{for } i \in (0, 1)$$

This is bias due to sample treated units who are not in the overlap region of the sample treated and population treated.


Example: Propensity Score Distributions
Sample Treated P(W)
Population Treated P(W)
150
100
50
0
-50
-100
-150
0
0.2
0.4
0.6
0.8
1
Propensity Score P(X)

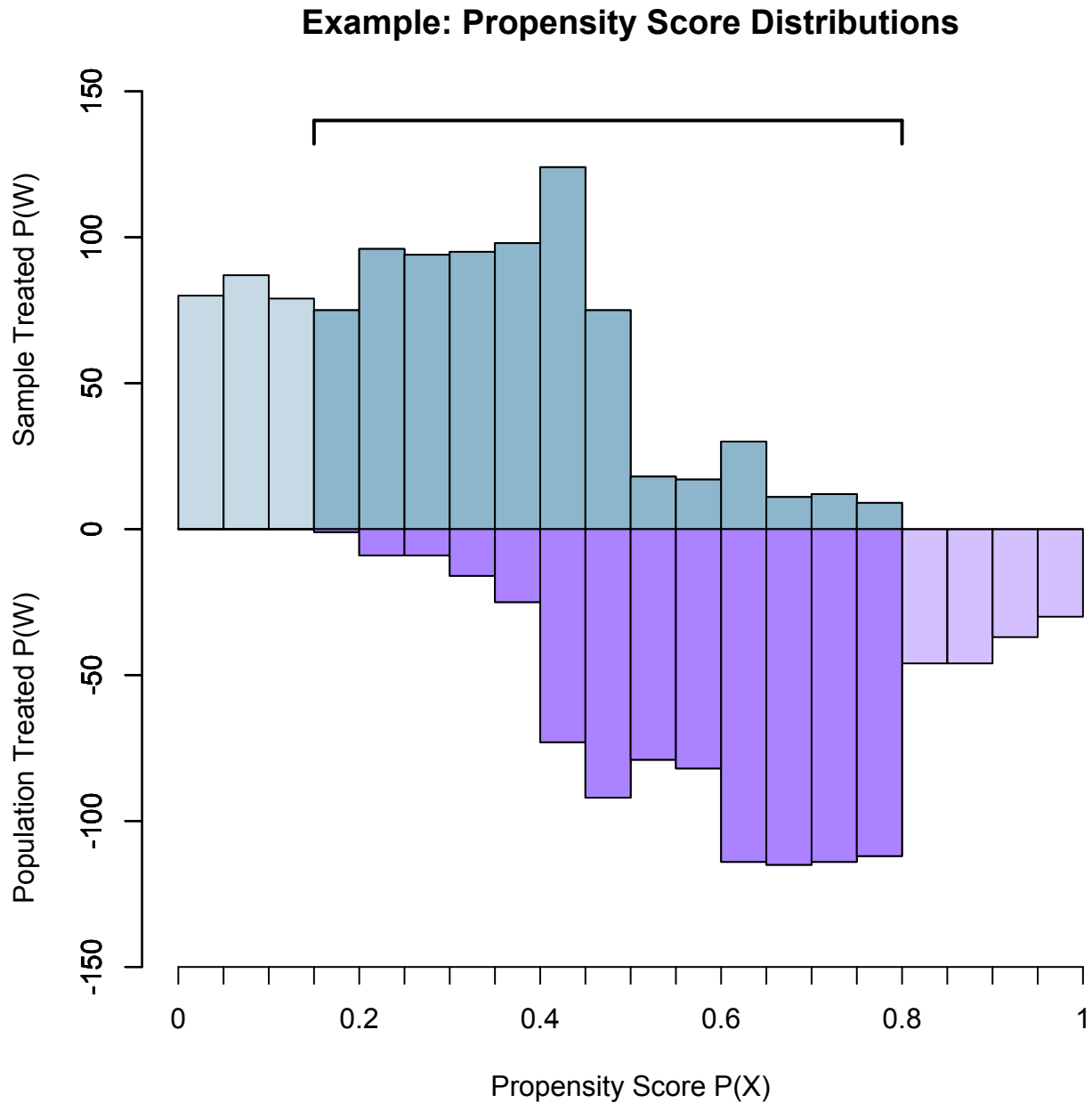
$$
\begin{aligned}
B_E = & \mathbb{E}_{S_{11} \setminus S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 1)\} \\
& - \mathbb{E}_{S_{01} \setminus S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 0)\} \\
& + \mathbb{E}_{dF(S=1) - dF(S=0), S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 0)\} \\
& + \mathbb{E}_{dF(S=1), S_{T1}} \{\mathbb{E}(Y_i | W, T = 1, S = 1) - \mathbb{E}(Y_i | W, T = 1, S = 0)\} \\
& \text{for } i \in (0, 1)
\end{aligned}
$$

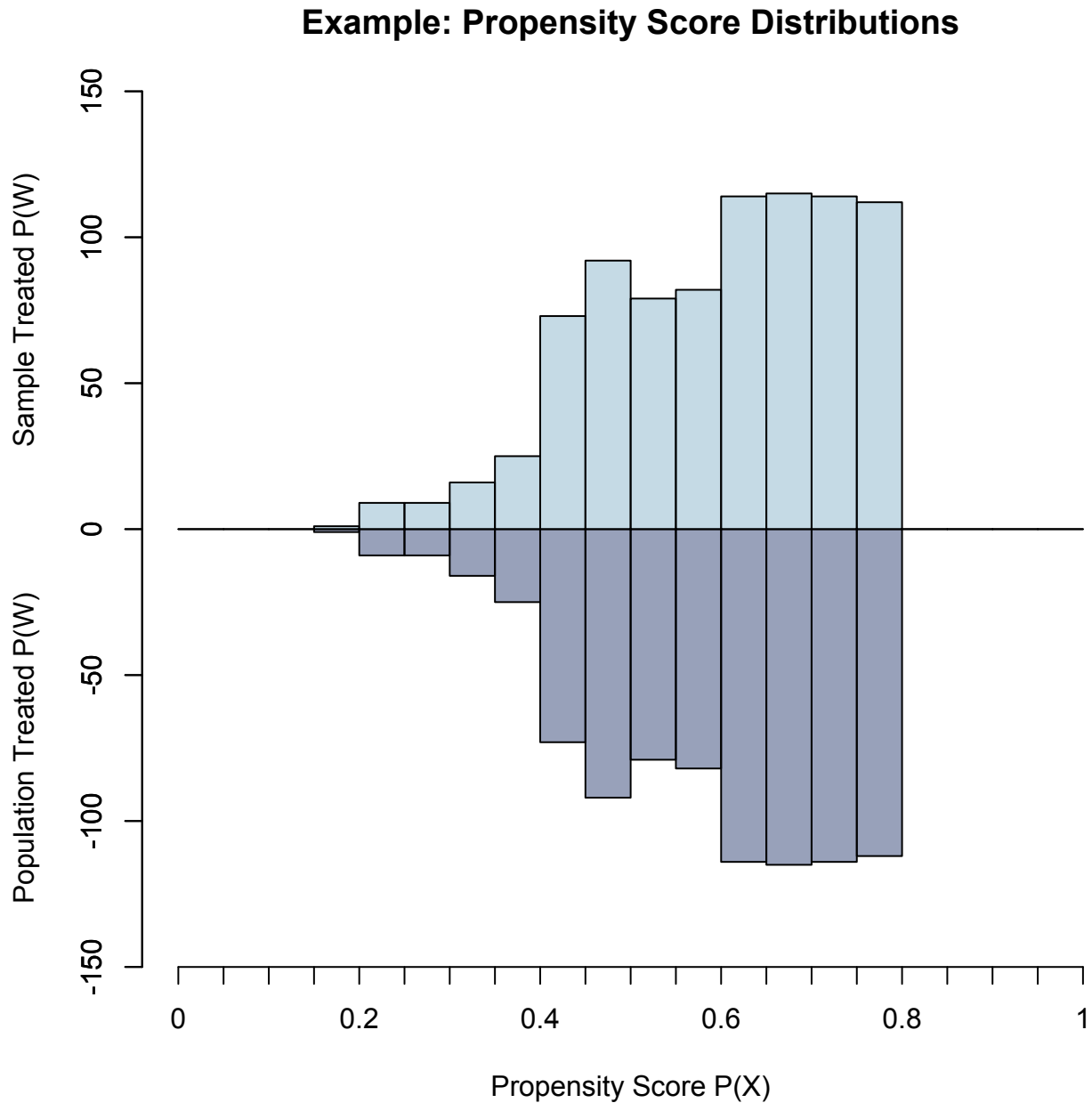
This is bias due to population treated units who are not in the overlap region of the sample treated and population treated.


Example: Propensity Score Distributions
Sample Treated P(W)
Population Treated P(W)
150
100
50
0
-50
-100
-150
0
0.2
0.4
0.6
0.8
1
Propensity Score P(X)

$$
\begin{aligned}
B_E = & \mathbb{E}_{S_{11} \setminus S_{T1}} \{\mathbb{E}(Y_i|W, T=1, S=1)\} \\
& - \mathbb{E}_{S_{01} \setminus S_{T1}} \{\mathbb{E}(Y_i|W, T=1, S=0)\} \\
& + \mathbb{E}_{dF(S=1)-dF(S=0), S_{T1}} \{\mathbb{E}(Y_i|W, T=1, S=0)\} \\
& + \mathbb{E}_{dF(S=1), S_{T1}} \{\mathbb{E}(Y_i|W, T=1, S=1) - \mathbb{E}(Y_i|W, T=1, S=0)\} \\
& \text{for } i \in (0,1)
\end{aligned}
$$

This is bias due to imbalance of the sample treated and population treated units in the overlap region.

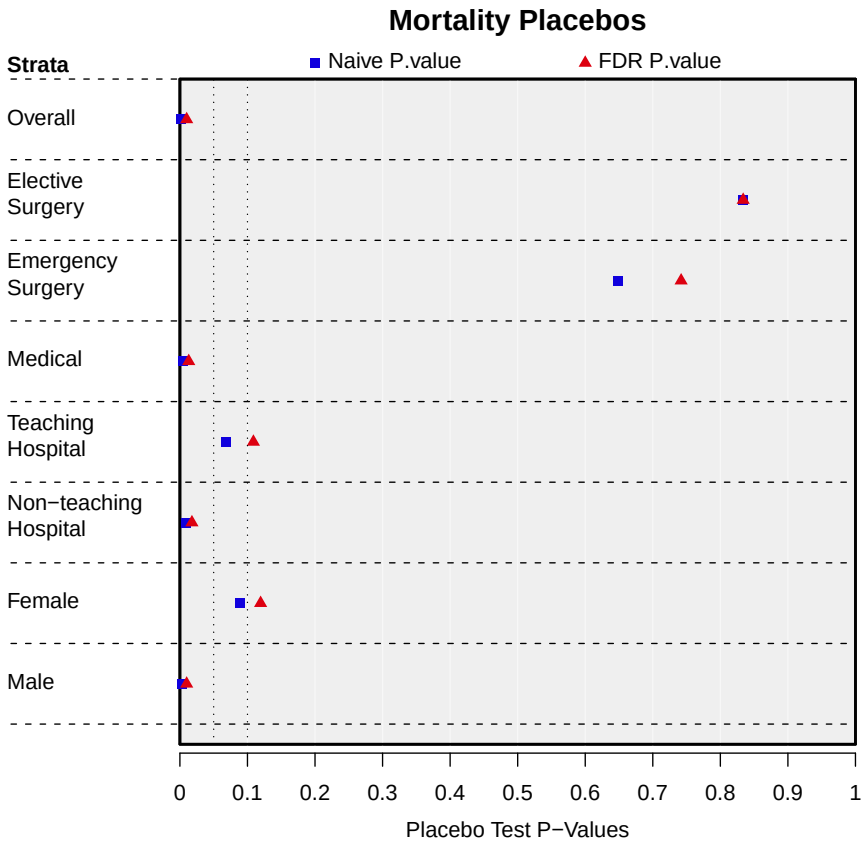

Example: Propensity Score Distributions
Sample Treated P(W)
Population Treated P(W)
150
100
50
0
-50
-100
-150
0
0.2
0.4
0.6
0.8
1
Propensity Score P(X)

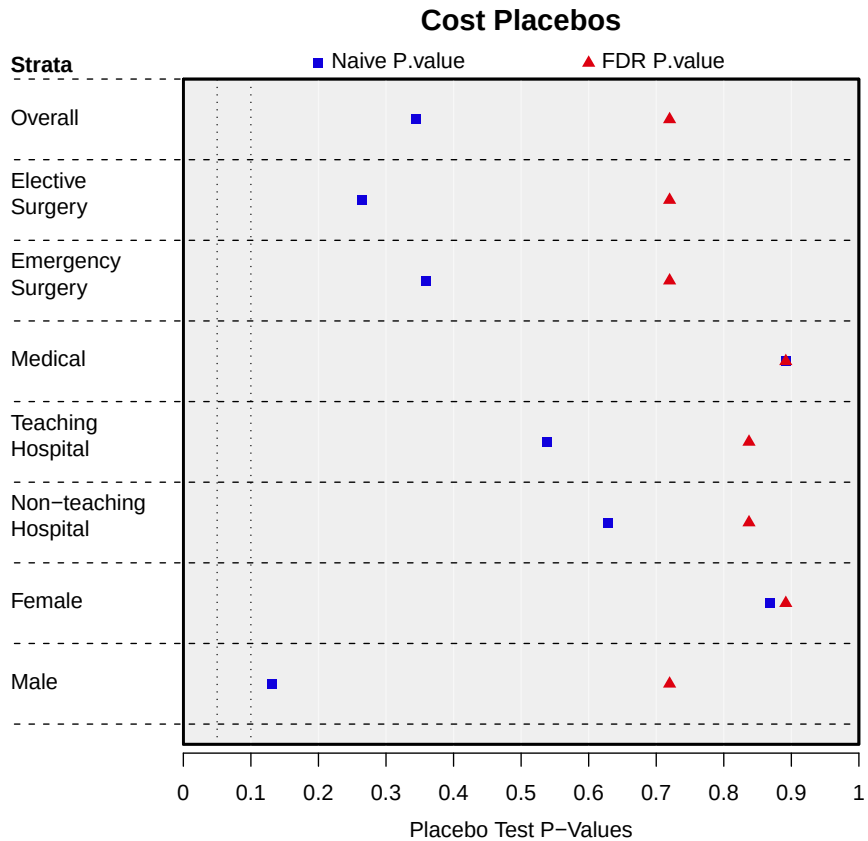

Example: Propensity Score Distributions
Sample Treated P(W)
Population Treated P(W)
150
100
50
0
-50
-100
-150
0
0.2
0.4
0.6
0.8
1
Propensity Score P(X)

$$
\begin{aligned}
B_E = & \mathbb{E}_{S_{11} \setminus S_{T1}} \{\mathbb{E}(Y_i|W, T=1, S=1)\} \\
& - \mathbb{E}_{S_{01} \setminus S_{T1}} \{\mathbb{E}(Y_i|W, T=1, S=0)\} \\
& + \mathbb{E}_{dF(S=1)-dF(S=0), S_{T1}} \{\mathbb{E}(Y_i|W, T=1, S=0)\} \\
& + \mathbb{E}_{dF(S=1), S_{T1}} \{\mathbb{E}(Y_i|W, T=1, S=1) - \mathbb{E}(Y_i|W, T=1, S=0)\} \\
& \text{for } i \in (0,1)
\end{aligned}
$$

This is the usual definition of bias, or the sample selection bias. There is nothing we can do to fix this, and the bias can be of opposite sign and any magnitude.

Figure: Alternative Approach: Inverse Propensity Score with Random Forest





Maximum Entropy

Jaynes defined the principle of maximum entropy as:

$$\max_{\mathbf{p}} S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i$$

$$s.t. \begin{cases} \sum_{i=1}^{n} p_i = 1 \\ \sum_{i=1}^{n} p_i g_r(x_i) = \sum_{i=1}^{n} p_i g_{ri} = a_r \qquad r = 1, \ldots, m \\ p_i \geq 0 \qquad i = 1, 2, \ldots, n \end{cases}$$

- Equation (1) is referred to as a natural constraint, stating that all probabilities must sum to unity.
- Equation (2), the m moment constraints, are referred to as the consistency constraints. Each a_r represents an r-th order moment, or characteristic moment, of the probability distribution.
- Equation (3) the final constraint ensures that all probabilities are non-negative. This is always met.

Genetic Matching (GenMatch)

Genetic matching is a method for performing multivariate matching. GenMatch:

- algorithmically maximizes the balance of observed potential confounders across matched treated and control units
- uses an evolutionary search algorithm to determine the weight each covariate is given
- Theorems in support of GAs are based on interpreting them as finite and irreducible Markov chains.

Properties of Matching Algorithms

- When can matching confounders make bias worse? e.g., what if the propensity score model is incorrect?
- All affinely invariant matching methods have the Equal Percent Bias Reduction (EPBR) property under some conditions.
- If X are distributed with ellipsoidal distributions, then the EPBR property holds for affinely invariant matching methods (Rubin and Thomas 1992).
- There is an extension to a restricted class of mixtures (Rubin and Stuart 2006): discriminant mixtures of proportional ellipsoidally symmetric distributions.

Equal Percent Bias Reduction (EPBR)

- Let Z be the expected value of X in the matched control group. Then we say that a matching procedure is EPBR if

$$E(X|T=1) - Z \quad = \gamma \{E(X|T=1) - E(X|T=0)\}$$

  for a scalar $0 \leq \gamma \leq 1$.
- We say that a matching method is EPBR for X because the percent reduction in the mean biases for each of the matching variables is the same.
- In general, if a matching method is not EPBR, then the bias for some linear function of X is increased.
- We may care about nonlinear functions of X.

Mahalanobis Distance

- The most common method of multivariate matching is based on the Mahalanobis distance. The Mahalanobis distance measure between any two column vectors is defined as:

$$md(X_i, X_j) = \left\{(X_i - X_j)' S^{-1} (X_i - X_j)\right\}^{\frac{1}{2}}$$

  where X_i and X_j are two different observations and S is the sample covariance matrix of X.
- Mahalanobis distance is an appropriate distance measure if each covariate has an elliptic distribution whose shape is common between treatment and control groups.
- In finite samples, Mahalanobis distance will not be optimal.

More General Method of Measuring Distance

- A more general way to measure distance is defined by:

$$d(X_i, X_j) = \left\{ (X_i - X_j)' \left(S^{-1/2}\right)' W S^{-1/2} (X_i - X_j) \right\}^{\frac{1}{2}}$$

  where W is a $k \times k$ positive definite weight matrix and $S^{1/2}$ is the Cholesky decomposition of S which is the variance-covariance matrix of X.
- All elements of W are zero except down the main diagonal. The main diagonal consists of k parameters which must be chosen.
- This leaves the problem of choosing the free elements of W. For identification, there are only $k - 1$ free parameters.

Optimization

- Many loss functions are possible. Such as:
  - minimize Kullback-Leiber divergence measure
  - minimize the largest discrepancy
  - minimize the mean or median discrepancy
  - minimize some other quantile
  - restrict the above to only uniformly improving moves
- All loss functions make an assumption about the response surface
- Why note use knowledge of Y, the outcome variable?

Genetic Optimization

- The optimization problem described above is difficult and irregular, and we utilize an evolutionary algorithm developed by Sekhon and Mebane 1998 called GENOUD.
- Random search also works better than the usual matching methods, but is less efficient than GENOUD.