Combining Randomized Control Trials and Observational Data in the Age of Big Data

Jasjeet S. Sekhon

UC Berkeley

The Opportunity

- Explosion of data sources: administrative, electronic medical records, online behavior
- Population data is becoming more common and precise
- How can it be used?
- Interest in fine-grained inference: e.g., subgroups, heterogeneous effects
- Some traditional experimental design methods have become computationally infeasible
- Researcher's degrees of freedom has increased
- Big rise in false positive rate

The Problem

- Randomized Controlled Trials (RCTs) are rare and often small, especially a problem with medical experiments
- RCTs usually not conducted on the population of interest
- Combine information from both RCTs and population data to estimate treatment effects in the population
- Precise targeting of treatments, e.g., precision medicine

The Sample Selection Problem

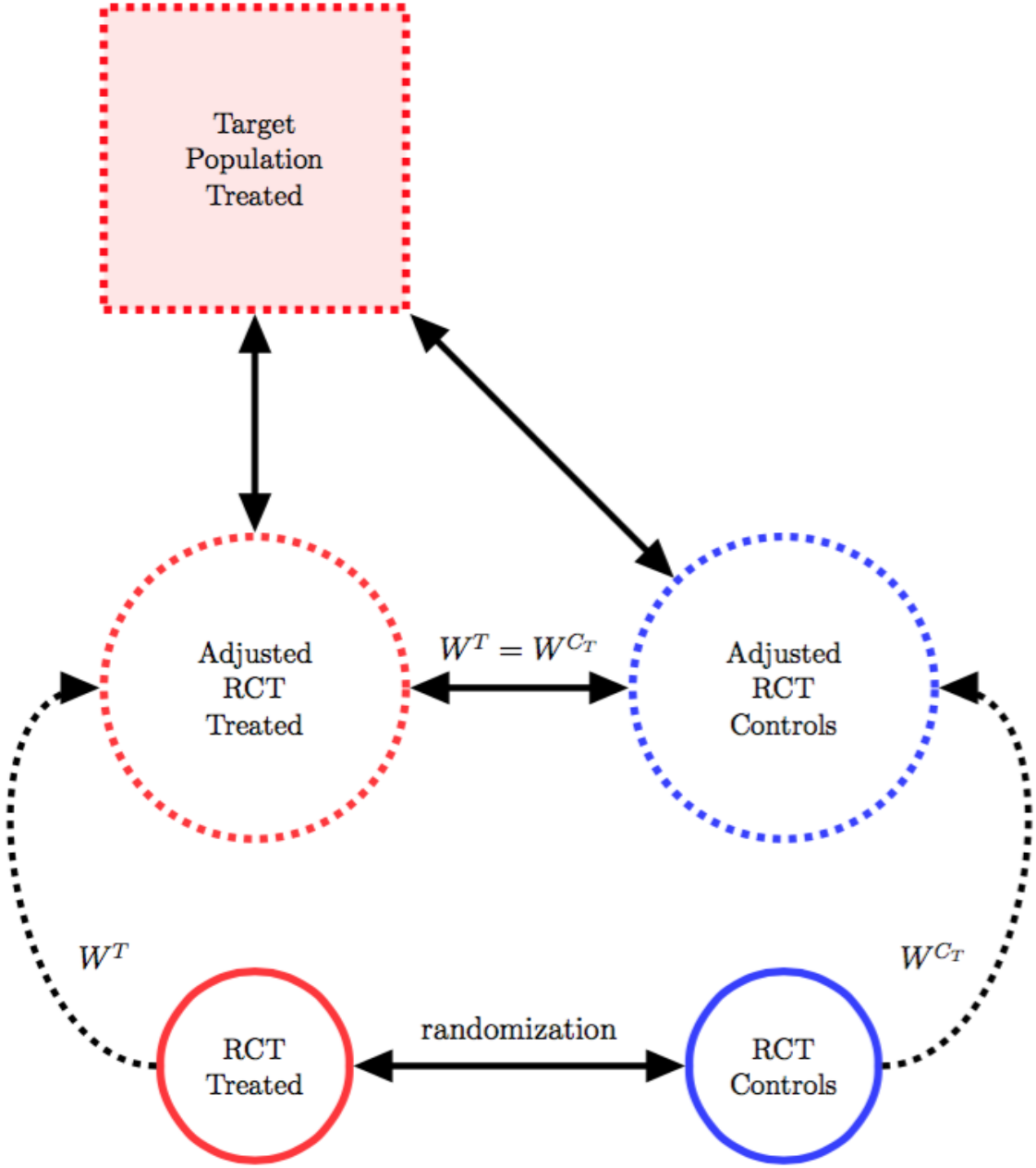
- We want to make inferences for the full population of interest:
  - RCTs raise issues of Randomization Bias (Heckman and Smith 1995): poor external validity
  - NRSs raise issues of Selection Bias, or non random assignment to treatment
- How to combine information from RCTs and NRSs?

Pulmonary Artery Catheterization (PAC)

- PAC is an invasive cardiac monitoring device for critical ill patients (ICU)—e.g., myocardial infarction (ischaemic heart disease)
- PAC-man trail n=1,013
- RCT find no effect; seven NRS find that PAC increases mortality (e.g., Connors et al. JAMA 1996)
- Registry data: 1.5million ICU admissions. 1,052 PAC cases and 31,447 potential controls

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 between the two studies differ, and we are interested in identifying PATT.


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

Neyman Model

- Fundamental problem: not observing all of the potential outcomes or counterfactuals
- Let Y_{i1} denote i's outcome when i is in the treatment regime
- Let Y_{i0} denote i's outcome when i is in the control regime
- Let T_i be a treatment indicator: 1 when i is in the treatment regime and 0 otherwise
- With no interference, the observed outcome for observation i is $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$
- The treatment effect for i is $\tau_i = Y_{i1} - Y_{i0}$

Experimental Data

- If assignment to treatment is randomized, the inference problem is straightforward because the two groups are from the same population: $\{Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i\}$.
- The Sample Average Treatment Affect (SATE) is simply:

$$\begin{aligned} \bar{\tau} &= \mathbb{E}(Y_1 - Y_0) \\ &= \mathbb{E}(Y|T=1) - \mathbb{E}(Y|T=0) \end{aligned}$$

Some Definitions

Take a sample of N units, $i = 1, 2, \ldots,$ from a large population

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

Estimands

Population Treatment Effects:

$$\begin{aligned}
\tau_{PATE} &= \mathbb{E}(Y_{01} - Y_{00}|S = 0) \\
\tau_{PATC} &= \mathbb{E}(Y_{01} - Y_{00}|S = 0, T = 0) \\
\tau_{PATT} &= \mathbb{E}(Y_{01} - Y_{00}|S = 0, T = 1)
\end{aligned}$$

Sample Treatment Effects:

$$\begin{aligned}
\tau_{SATE} &= \mathbb{E}(Y_{11} - Y_{10}|S = 1) \\
\tau_{SAT*} &= \mathbb{E}(Y_{11}|S = 1, T = t) - \mathbb{E}(Y_{10}|S = 1, T = t),
\end{aligned}$$

where $t = 0$ for τ_{SATC} and $t = 1$ for τ_{SATT}

Consistency and SUTVA

Let us assume:

A.1 Treatment is consistent across studies:

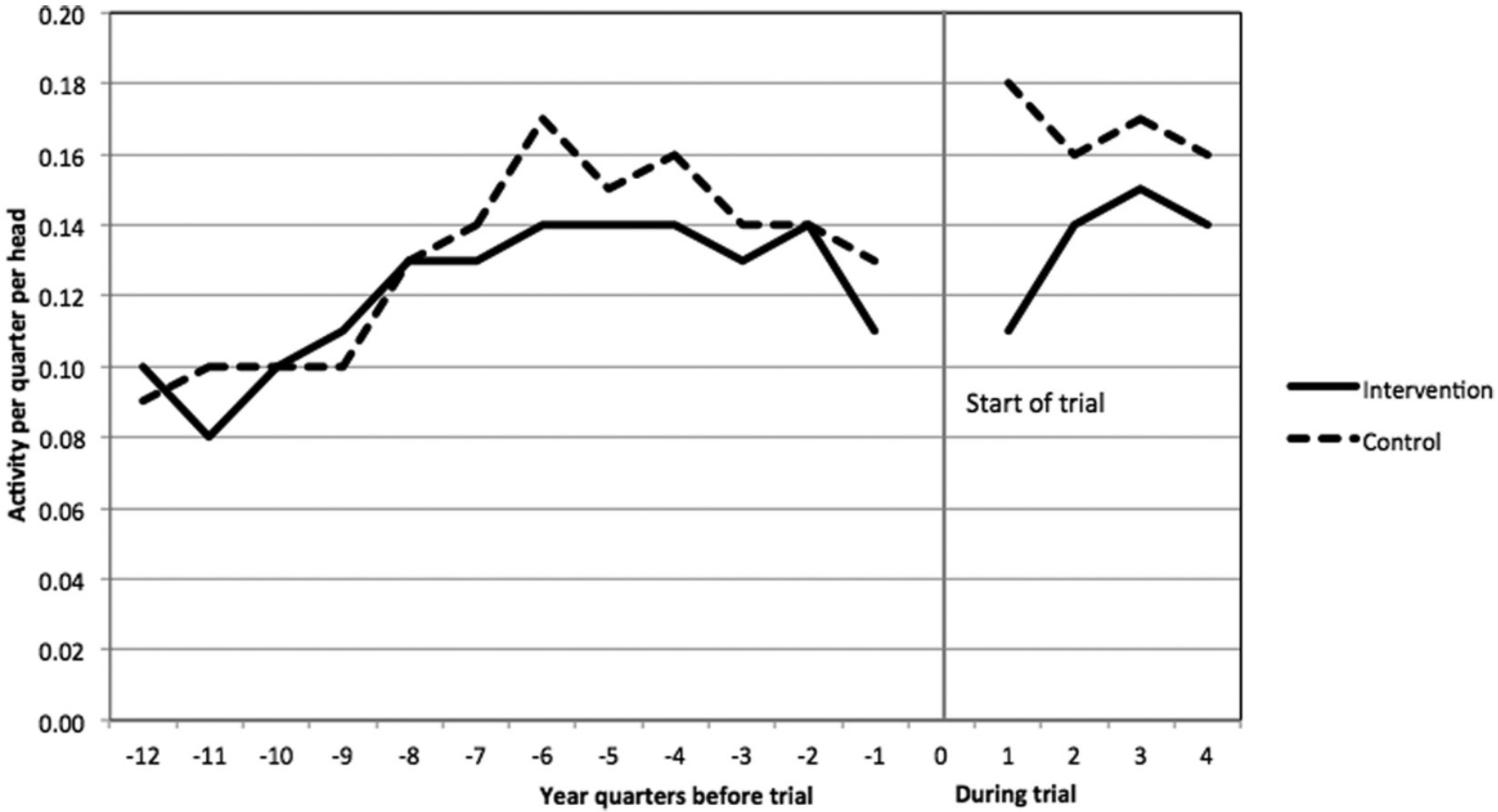
- **a** $Y_{i00} = Y_{i10}$
- **b** $Y_{i01} = Y_{i11}$

A.4 SUTVA: no interference between units

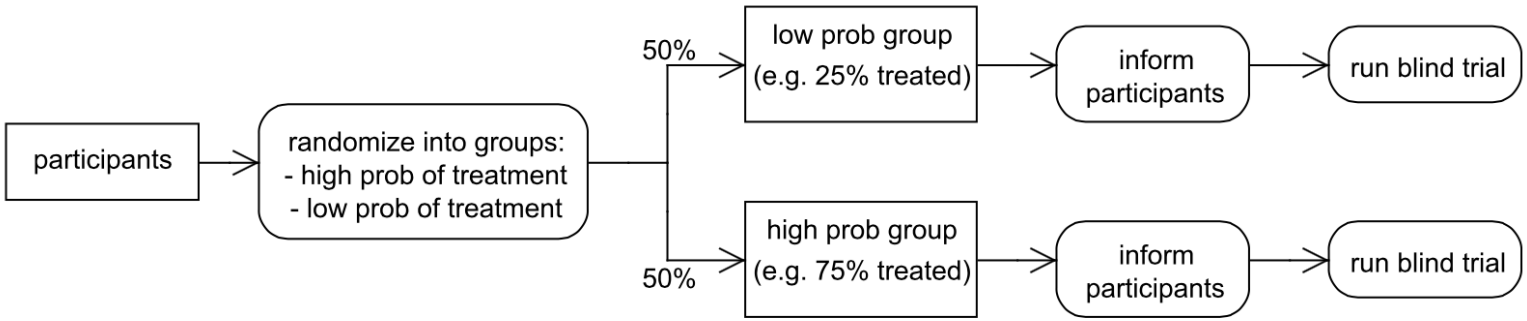
Then, we may write the potential outcomes for unit i as simply Y_0, Y_1, since t=0 or 1

These are not innocuous assumptions

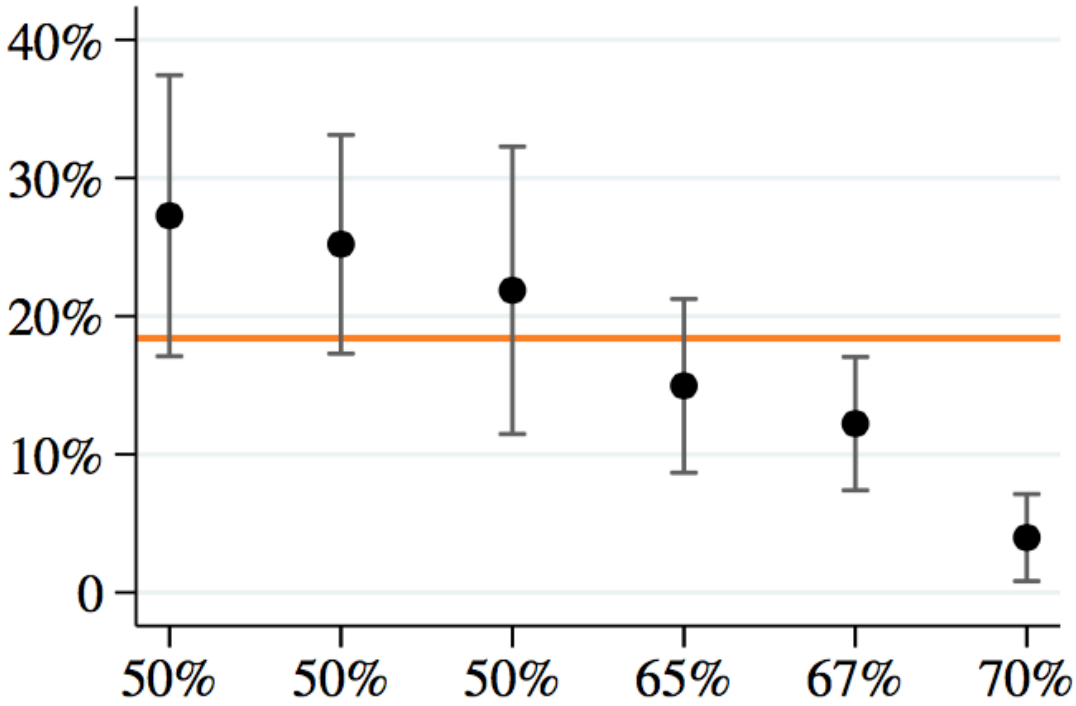
TeleHealth


0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Activity per quarter per head
-12
-11
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
Year quarters before trial
During trial
Start of trial
Intervention
Control

A Two-by-Two Blind Trial



Dropout Rates by Treatment Probability



Ignobility of Sample Assignment for Treated

A.2 Strong Ignorability of Sample Assignment for Treated:

$$Y_1 \perp\!\!\!\perp S|(W, T = 1) \; 0 < Pr(S = 1|W, T = 1) < 1$$

Under A.2, we can identify:

$$\begin{aligned} E(Y_1|S = 0, T = 1) &= E\{E(Y_1|S = 0, T = 1, W)|S = 0, T = 1\} \\ &= E\{E(Y_1|S = 1, T = 1, W)|S = 0, T = 1\} \\ &= E\{E(Y|S = 1, T = 1, W)|S = 0, T = 1\} \end{aligned}$$

This implies the following placebo test:

$$E(Y|S = 0, T = 1) = E\{E(Y|S = 1, T = 1, W)|S = 0, T = 1\}$$

Ignobility of Sample Assignment for Treated-Controls

A.3 Strong Ignorability of Sample Assignment for Treated-Controls:

$$Y_0 \perp\!\!\!\perp S|(W, T = 1) \; 0 < Pr(S = 1|W, T = 1) < 1$$

Under A.3 and randomization in the RCT, $Y_0 \perp\!\!\!\perp T|S = 1$:

$$\begin{aligned} E(Y_0|S = 0, T = 1) &= E\{E(Y_0|S = 0, T = 1, W)|S = 0, T = 1\} \\ &= E\{E(Y_0|S = 1, T = 1, W)|S = 0, T = 1\} \\ &= E\{E(Y_0|S = 1, T = 0, W)|S = 0, T = 1\} \\ &= E\{E(Y|S = 1, T = 0, W)|S = 0, T = 1\} \end{aligned}$$

PATT Identification

Therefore,

$$\begin{aligned}\tau_{PATT} &= E(Y_1|S=0,T=1) - E(Y_0|S=0,T=1) \\ &= E\left\{E(Y|S=1,T=1,W)|S=0,T=1\right\} \\ &\quad -E\left\{E(Y|S=1,T=0,W)|S=0,T=1\right\}\end{aligned}$$

But we could use the following, without A1b and A2:

$$\begin{aligned}\tau'_{PATT} &= E(Y_1|S=0,1=1) - E(Y_0|S=0,T=1) \\ &= E(Y|S=0,T=1) \\ &\quad -E\left\{E(Y|S=1,T=0,W)|S=0,T=1\right\},\end{aligned}$$

but no placebo test and break randomization

Remarks

- Even if A1, A2, and A3 are false, if we don't break randomization, we have an (asymptotically) unbiased estimator of the reweighted RCT
- We also break randomization if we condition on post-treatment variables, as some advocate
- The data can provide evidence against us

Estimation

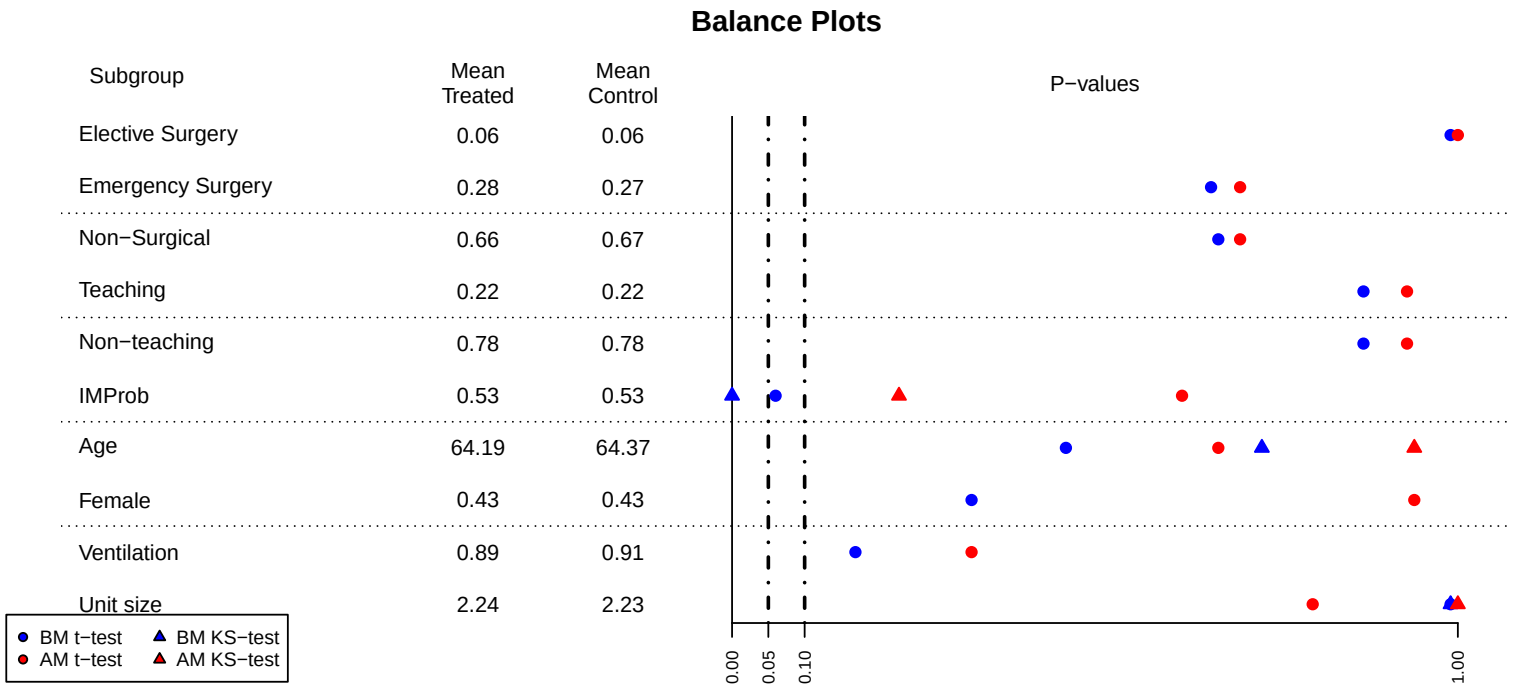
GenMatch:

- Matching method with automated balance optimization
- One would simply stratify if we had more data
- Testing/CIs estimation issues. Matching cannot use the bootstrap. Subsampling used

Maximum Entropy weighting:

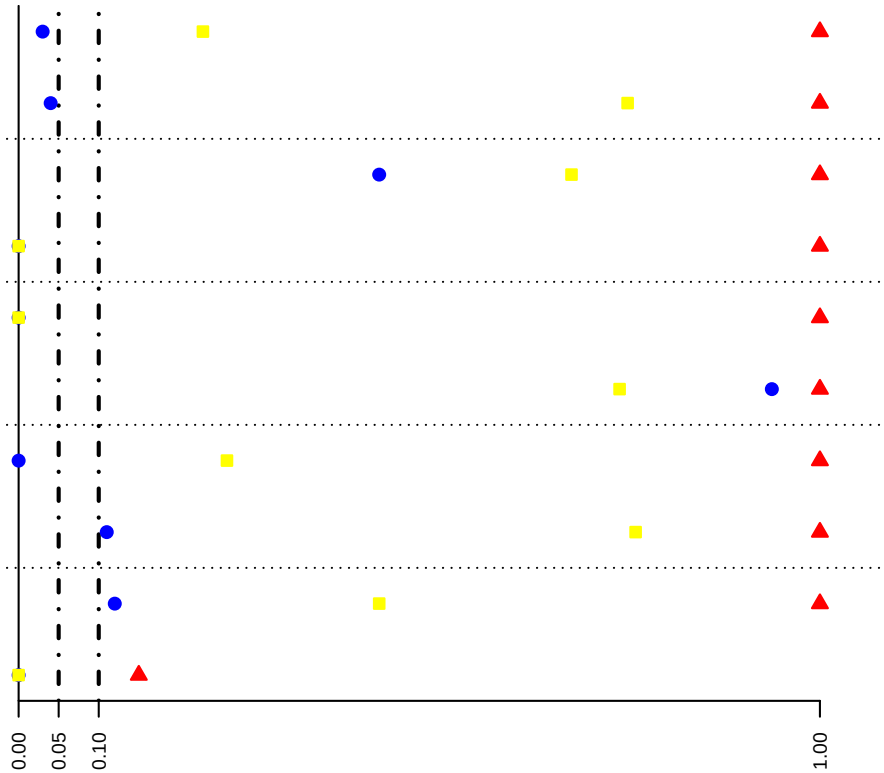
- Weighting method that assigns weights such that they simultaneously meet a set of consistency constraints while maximizing Shannon's measure of entropy
- Consistency constraints are based on moments of the population based on the NRS

Covariate Balance in RCT



Balance Before and After Adjustment

Subgroup	Mean RCT	Mean NRS	Mean RCT After Maxent Adjustment	Mean RCT After IPSW Adjustment
Elective Surgery	0.06	0.09	0.09	0.07
Emergency Surgery	0.28	0.23	0.23	0.24
Non−Surgical	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	63.49
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

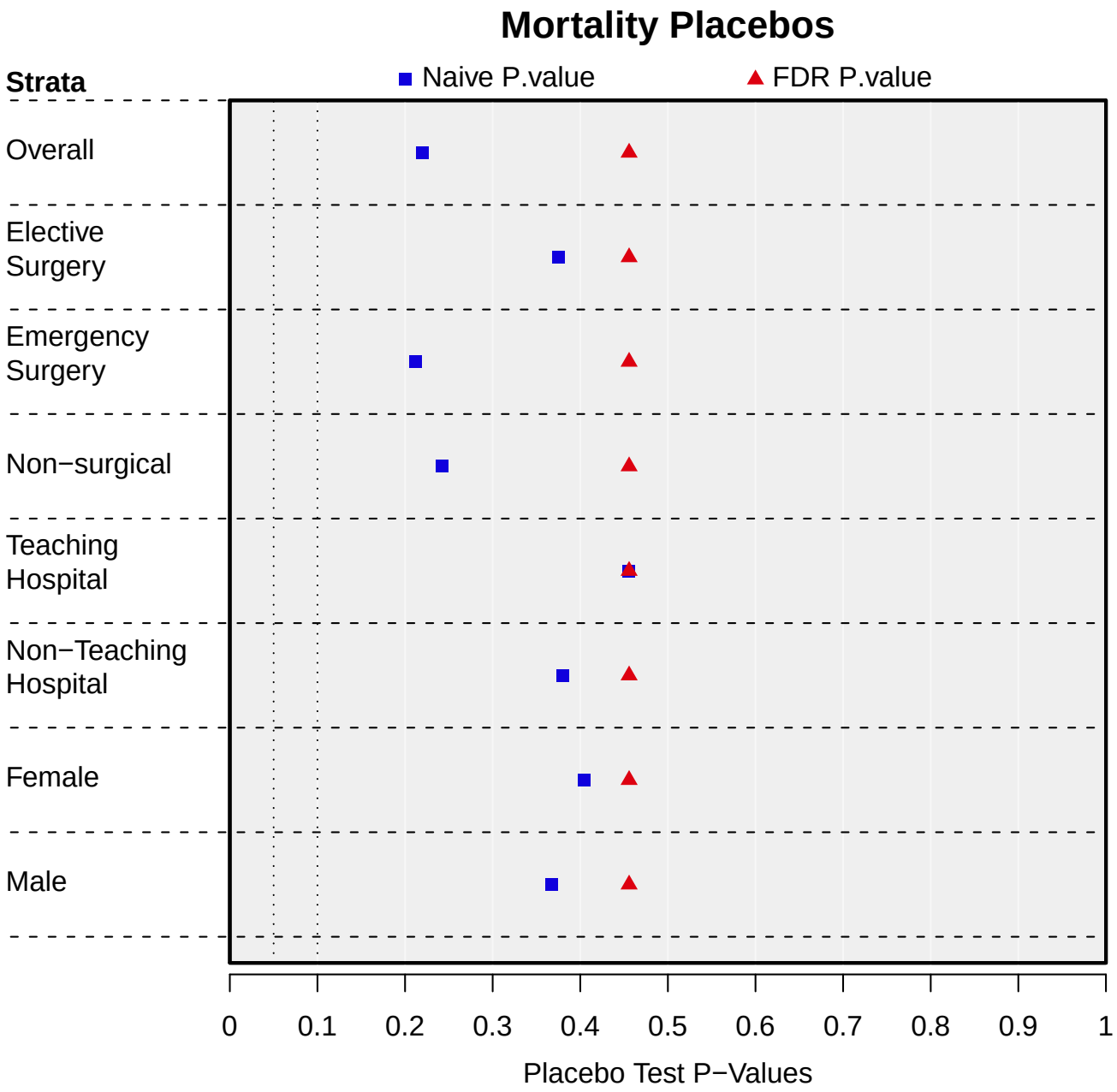


Placebo Test

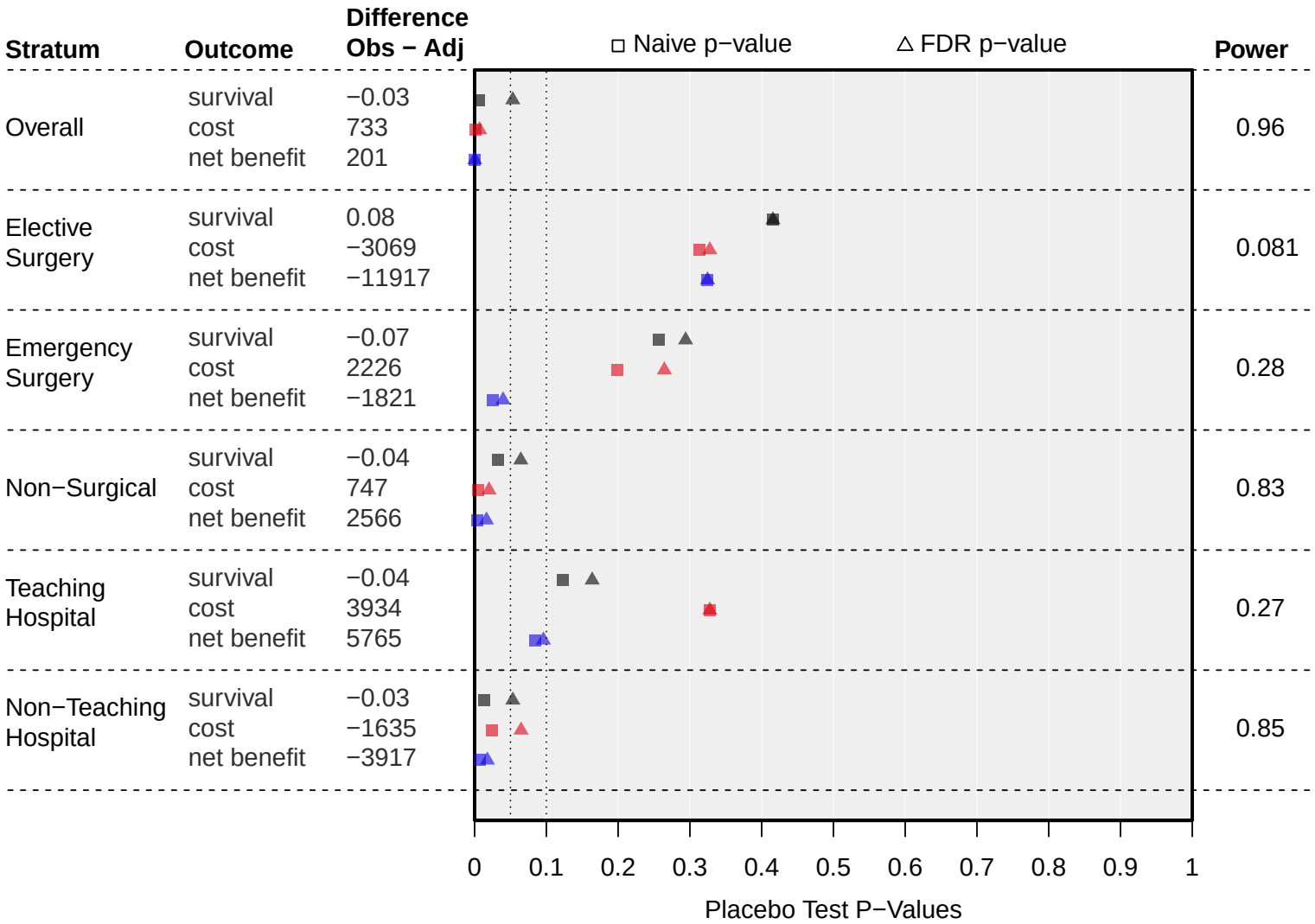
$$E(Y|S=0,T=1) \quad = \quad E\left\{E(Y|S=1,T=1,W)|S=0,T=1\right\}$$

- The difference between the mean outcome of the NRS treated and mean outcome of the reweighted RCT treated should be zero
- If not 0, at least one 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
- If both placebos are possible, one has assumed non-confounding in the NRS

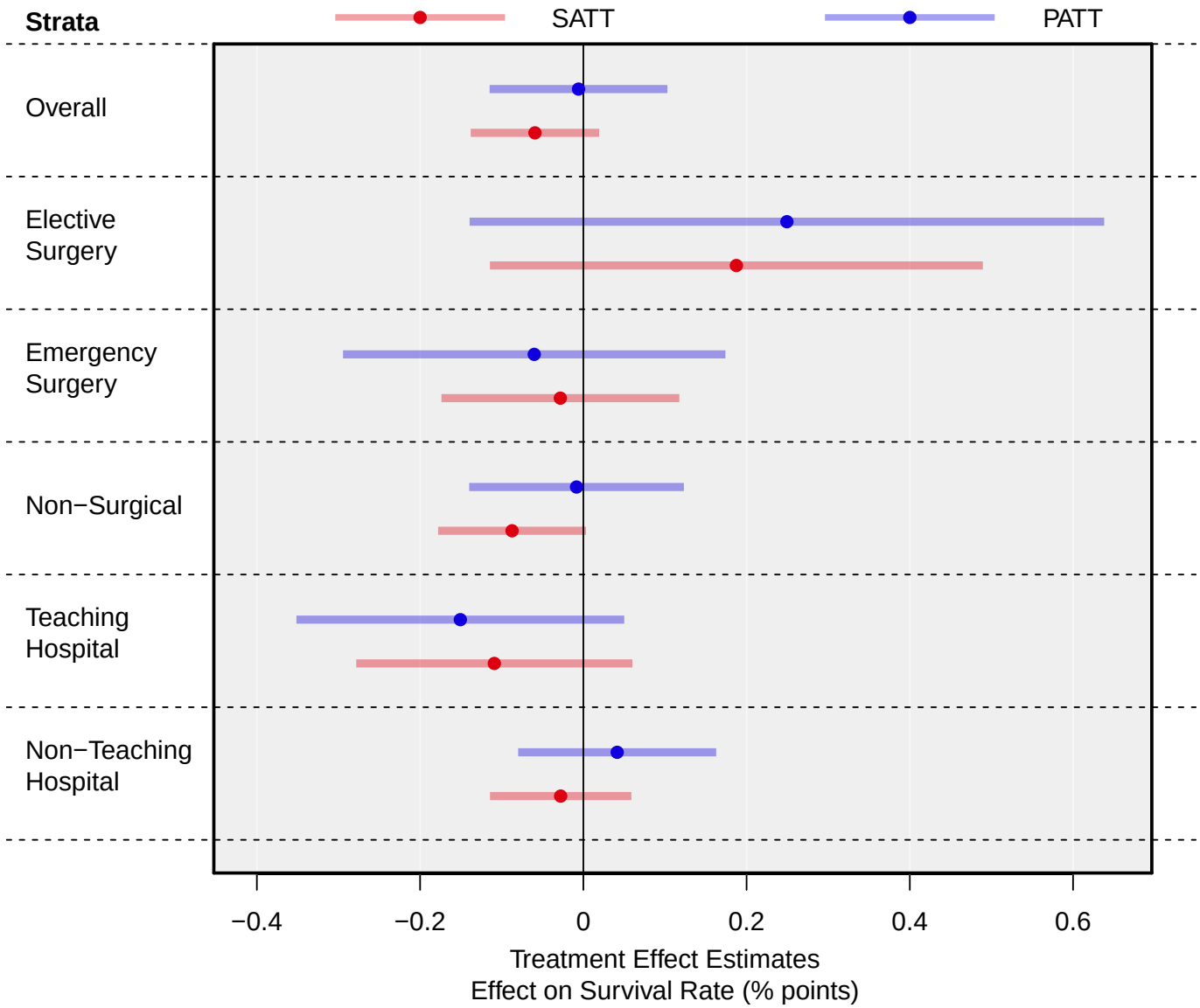
Placebo Tests (t-tests)



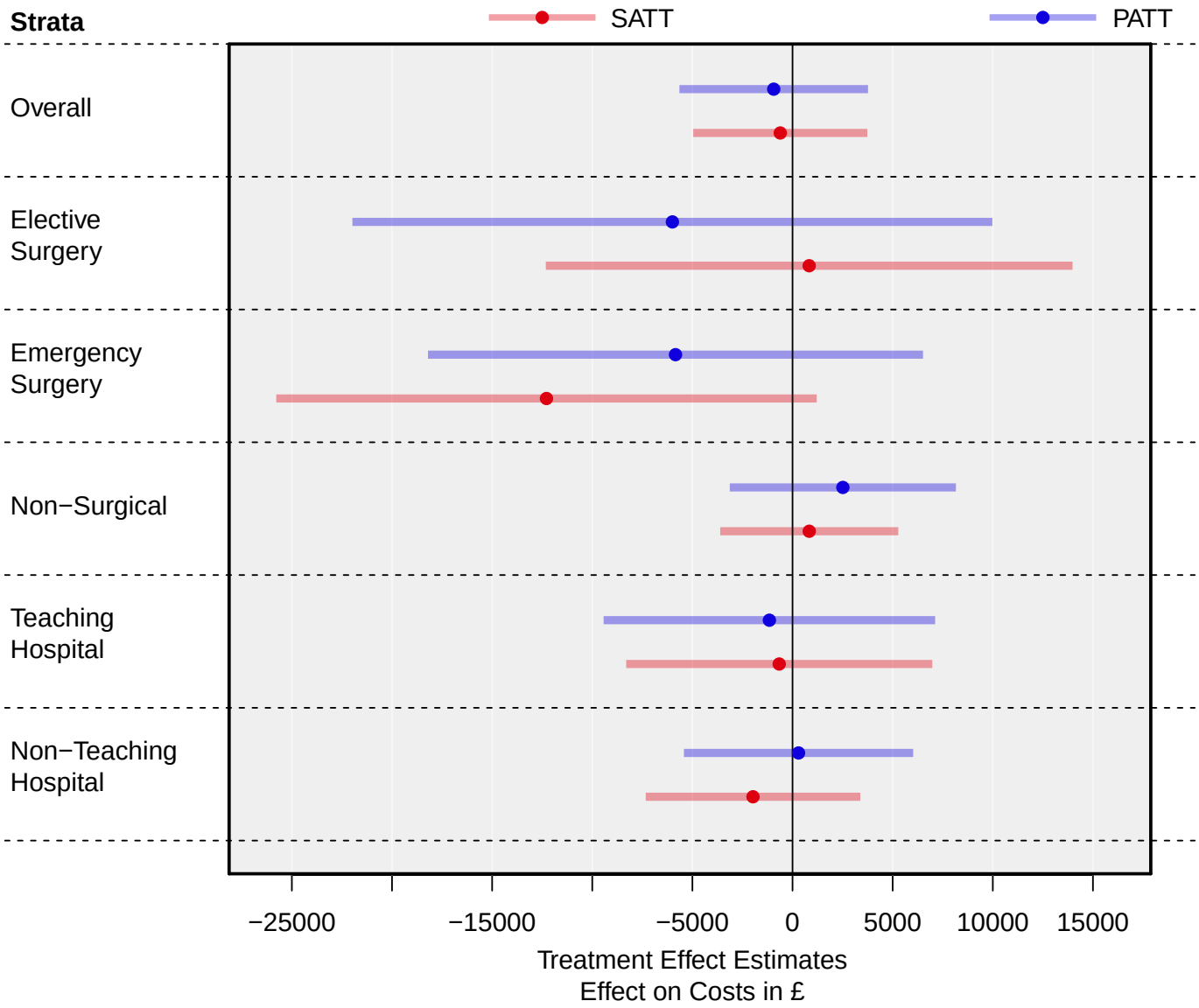
Placebo Tests (equivalence tests)



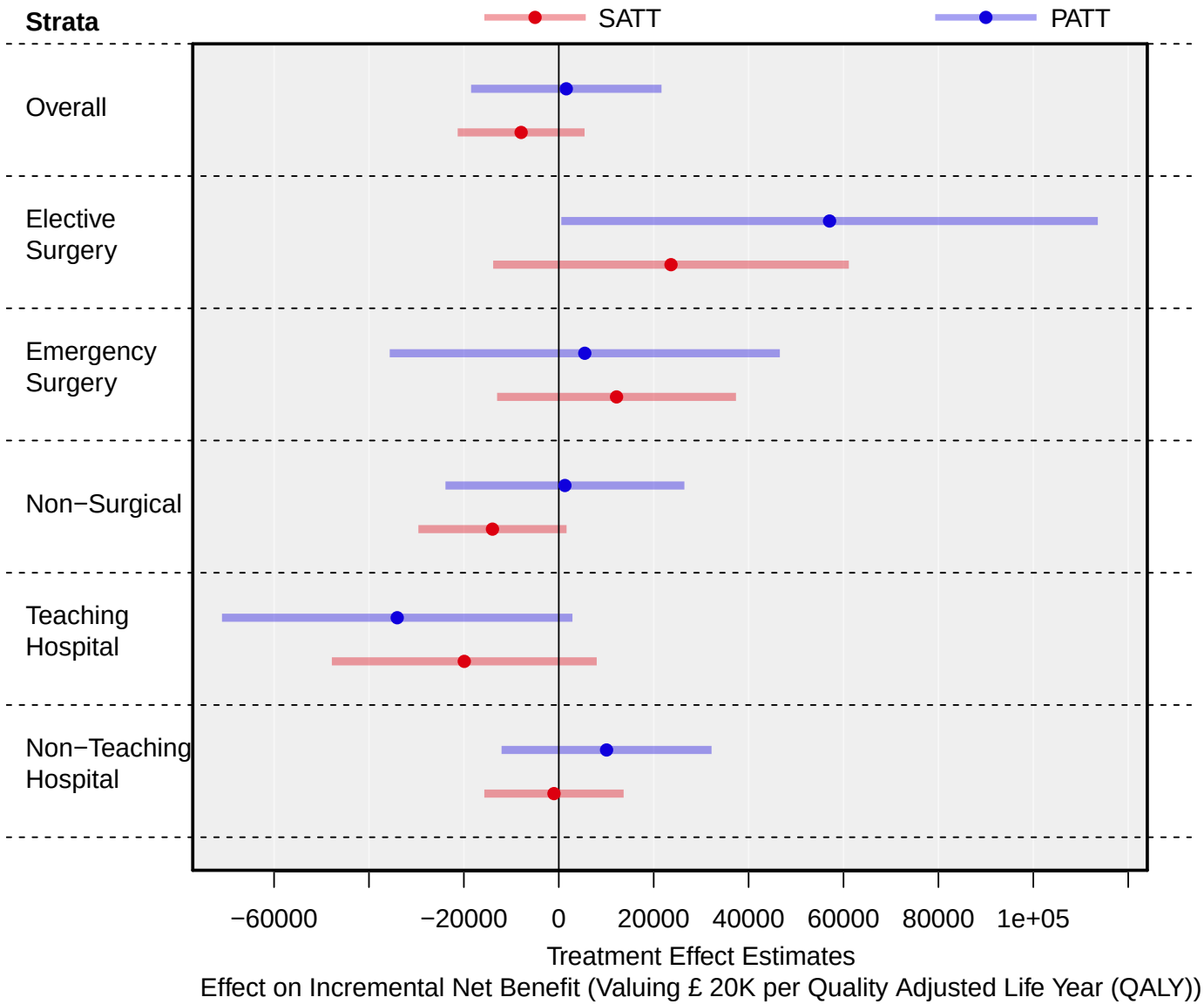
Population Treatment Effects on Hospital Survival Rates



Population Treatment Effects on Costs



Population Treatment Effects on Cost-Effectiveness



Why We Randomize?

- Unbiased estimator by design
- Make probability statements; "reasoned basis for inference" (Fisher, Peirce)
- Separate design from analysis (Cochran, Rubin)

A New Blocking Method

A new blocking method with theoretical properties

- Blocking: create strata and then randomize within strata
- Some analytical benefits for blocking, but the main one is transparency and minimizing fishing

A New Blocking Method

The method minimizes the pair-wise Maximum Within-Block Distance: λ

- Any valid distance metric; triangle inequality
- We prove this is a NP-hard problem
- Ensures good covariate balance by design: approximately optimal: $\leq 4 \times \lambda$
- Works for any number of treatments and any minimum number of observations per block
- It is fast: $O(n \log n)$ expected time
- It is memory efficient: $O(n)$ storage
- Special cases
  1. with one covariate: λ
  2. with two covariates: $\leq 2 \times \lambda$

Covariate imbalance in randomized experiments

- PROBLEM: In finite samples, there is a probability of bad covariate balance between treatment groups
- Bad imbalance on important covariates:
  - $\rightarrow$ Imprecise estimates of treatment effects
  - $\rightarrow$ Conditional bias
- In large samples problems remain: we want to estimate treatment effects for subgroups

Some theoretical results about blocking

- Blocking cannot hurt the precision of the estimator:
  - if no worse than random matching
  - if sample from an infinite super population

- Blocking may increase the estimated variance. But this is specific to the estimator used (degrees of freedom). e.g., randomization inference solves the problem.

Adjustment and covariate imbalance

- Regression adjustment [Freedman, 2008, Lin, 2012]
- Post-stratification [Miratrix, Sekhon, and Yu, 2013]:
  - Group similar units together after *after* randomization
  - SATE/PATE results good; *ex post* problems arise
  - Data mining concerns
- Re-randomization [Morgan and Rubin, 2012]:
  - Repeat randomly assigning treatments until covariate balance is "acceptable"
- LESSON: design the randomization to build in adjustment

Some Current blocking approaches

- Optimal Multivariate Matching Before Randomization [Greevy, Lu, Silber, and Rosenbaum, 2004]
- Matched-pairs blocking: Pair "most-similar" units together. For each pair, randomly assign one unit to treatment, one to control
- Optimal-greedy blocking [e.g. Moore, 2012]
- Some methods make principled probability statements impossible

Matched-Pairs

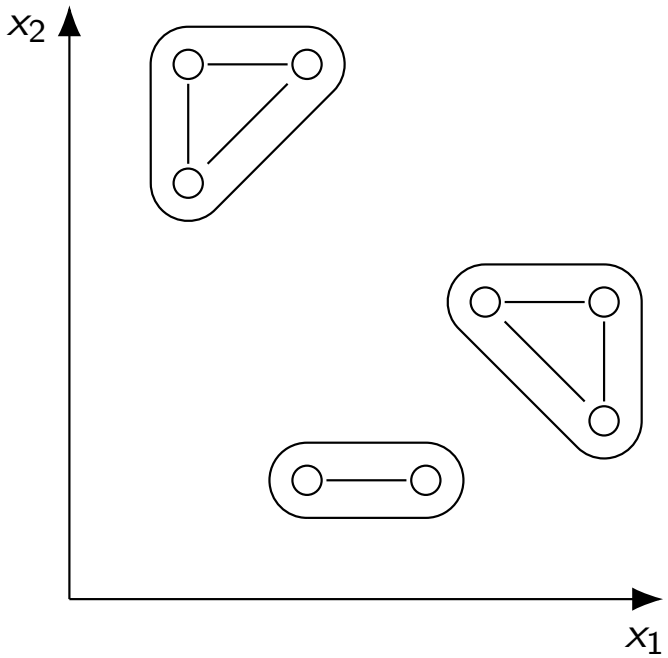
- No efficient way to extend approach to more than two treatment categories
- Fixed block sizes (2 units): design may pair units from different clusters
- Cannot estimate conditional variances [Imbens, 2011]
- Difficulty with treatment effect heterogeneity

Blocking by minimizing the Maximum Within-Block Distance (MWBD)

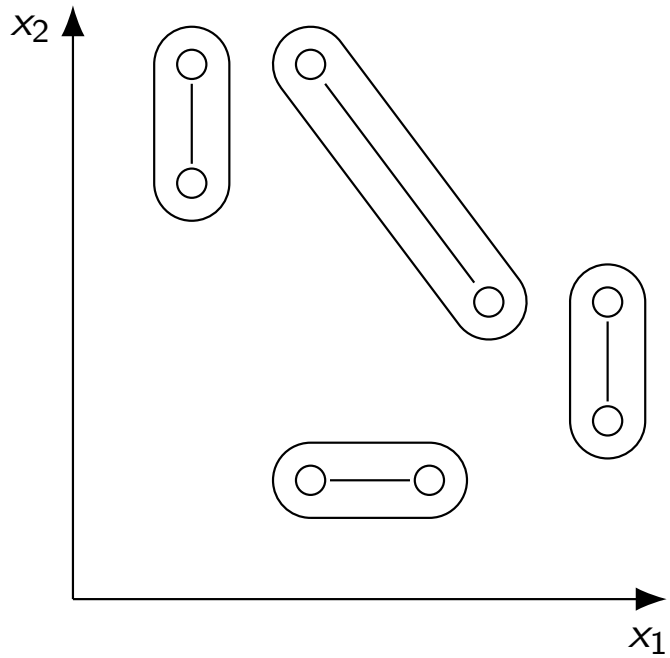
- Experiment with n units and t treatment categories
- Select a threshold $k \geq t$ for a minimum number of units to be contained in a block
- Block units so that each block contains at least k units, and so that the maximum distance between any two units within a block—the MWBD—is minimized
- Threshold k: Allows designs with multiple treatment categories, multiple replications of treatments within a block

Threshold blocking: relaxing the block structure

Threshold blocking



Fixed-sized blocking



An Advantage

Theorem

For all samples, all objective functions and all desired block sizes, the optimal threshold blocking is always weakly better than the optimal fixed-sized blocking.

- Proof: interpret blocking as an non-linear integer programming problem.
  - The search set of threshold blocking is a superset of fixed-sized blocking.

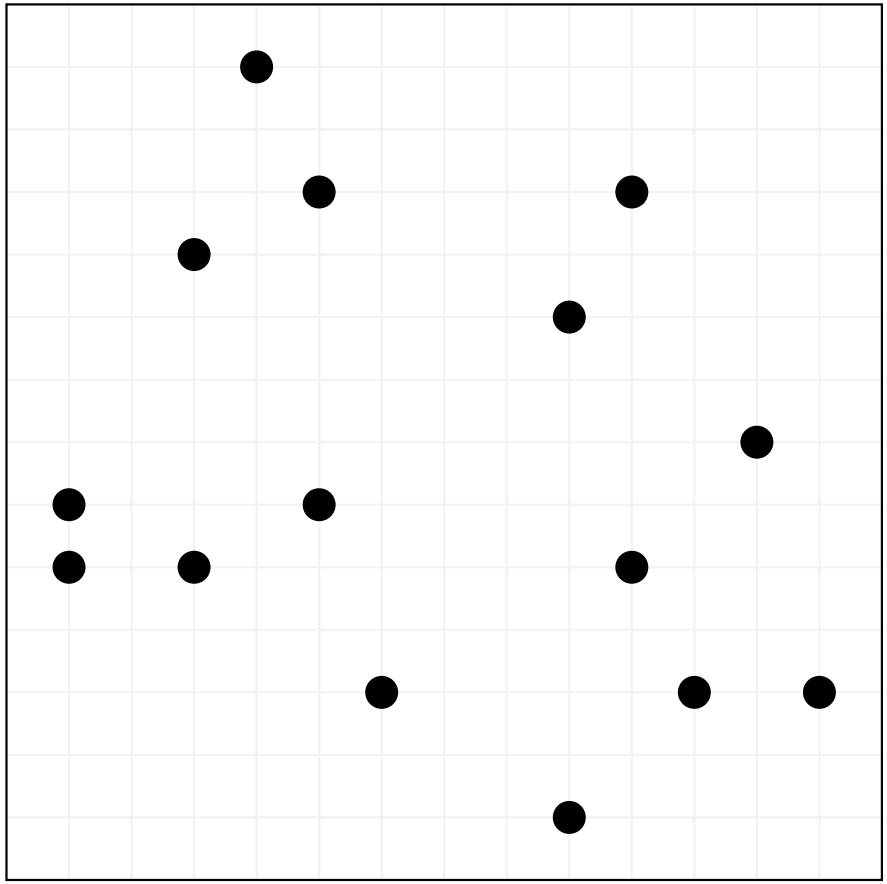
The AppOpt algorithm

Input:

- Units' covariates
- Distance metric
- Minimum block size: $k = 2$

Procedure:

1. A undirected complete graph with distances as edge weights
2. Find $(k-1)$-nearest neighbor graph
3. Construct the second power of NNG
4. Find a maximal independent set (seeds)
5. Form blocks with the seeds and their neighbors in NNG
6. Assign remaining units to a block containing any neighbor



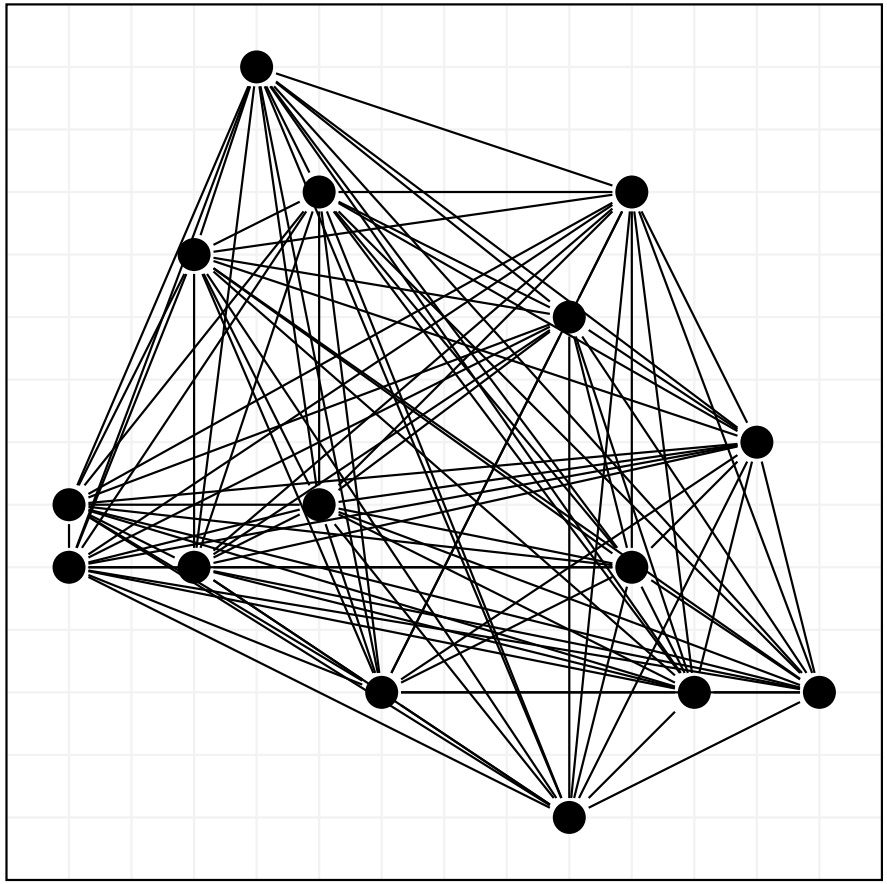
The AppOpt algorithm

Input:

- Units' covariates
- Distance metric
- Minimum block size: $k = 2$

Procedure:

1. A undirected complete graph with distances as edge weights
2. Find $(k-1)$-nearest neighbor graph
3. Construct the second power of NNG
4. Find a maximal independent set (seeds)
5. Form blocks with the seeds and their neighbors in NNG
6. Assign remaining units to a block containing any neighbor



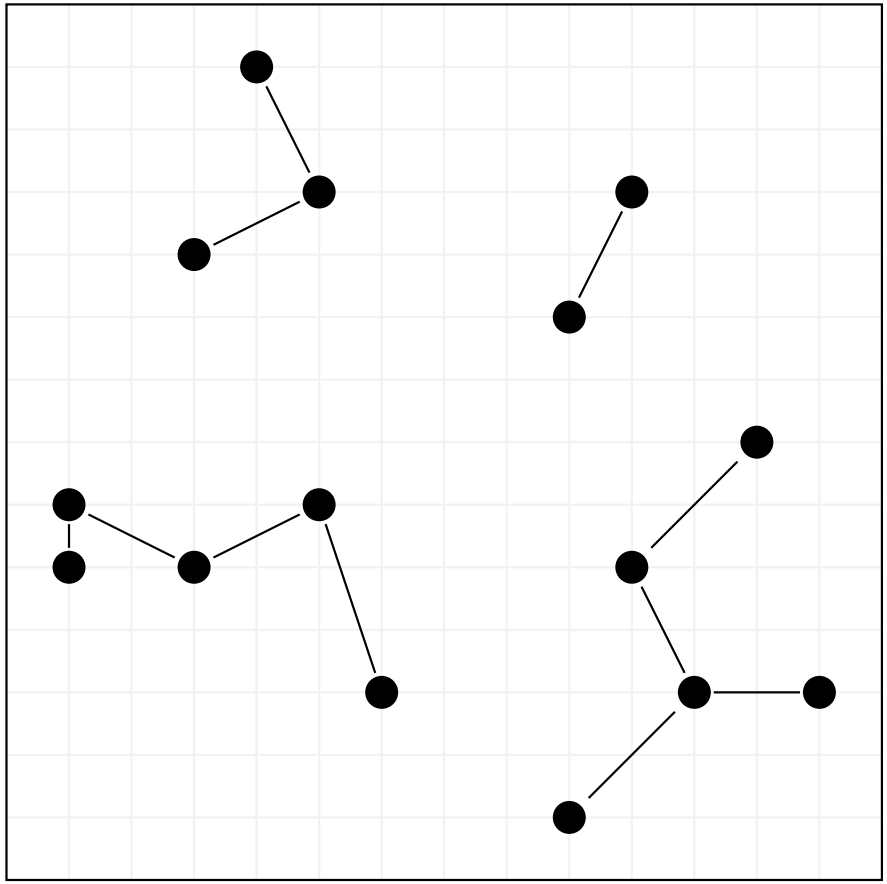
The AppOpt algorithm

Input:

- Units' covariates
- Distance metric
- Minimum block size: $k = 2$

Procedure:

1. A undirected complete graph with distances as edge weights
2. Find $(k-1)$-nearest neighbor graph
3. Construct the second power of NNG
4. Find a maximal independent set (seeds)
5. Form blocks with the seeds and their neighbors in NNG
6. Assign remaining units to a block containing any neighbor



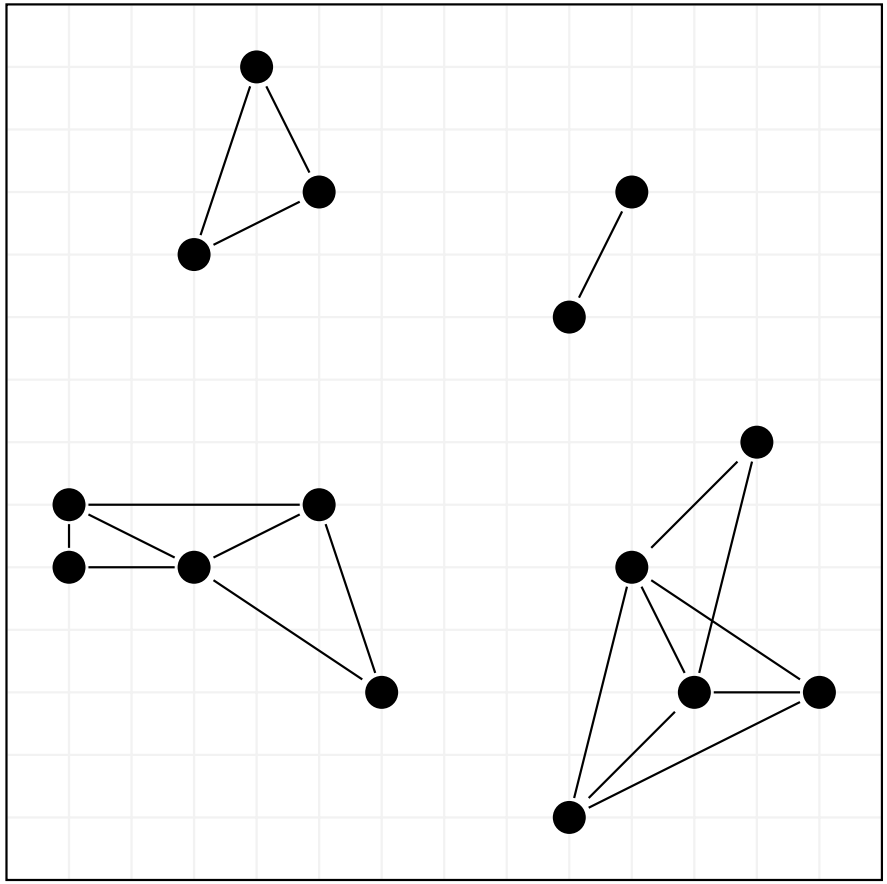
The AppOpt algorithm

Input:

- Units' covariates
- Distance metric
- Minimum block size: $k = 2$

Procedure:

1. A undirected complete graph with distances as edge weights
2. Find $(k-1)$-nearest neighbor graph
3. Construct the second power of NNG
4. Find a maximal independent set (seeds)
5. Form blocks with the seeds and their neighbors in NNG
6. Assign remaining units to a block containing any neighbor



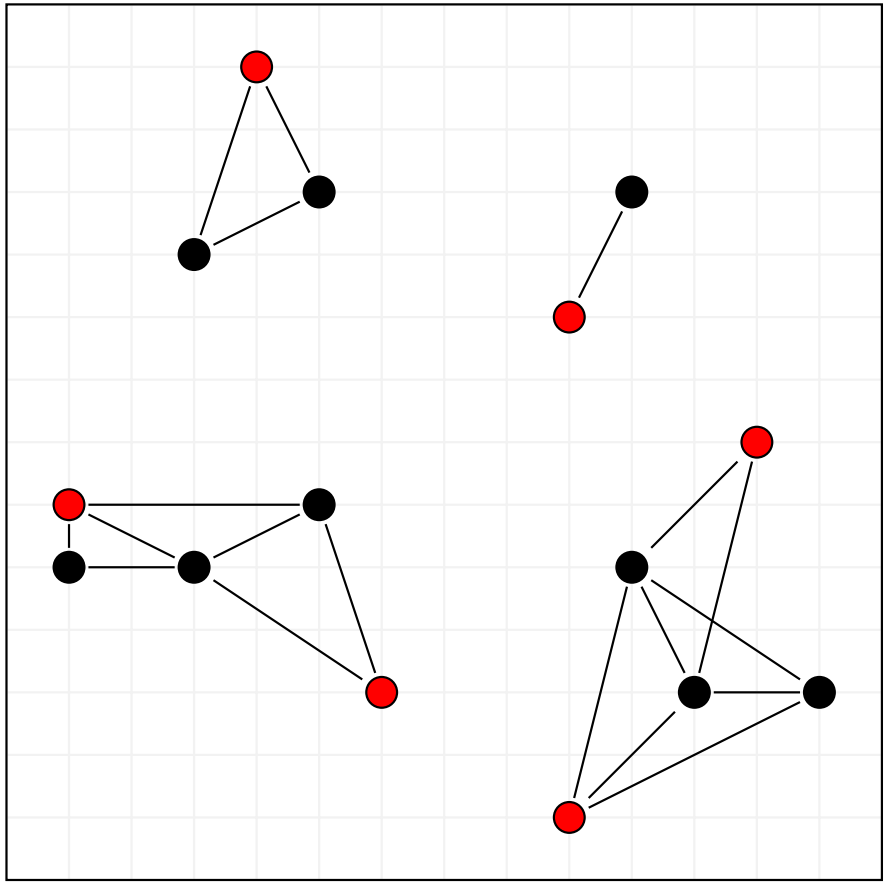
The AppOpt algorithm

Input:

- Units' covariates
- Distance metric
- Minimum block size: $k = 2$

Procedure:

1. A undirected complete graph with distances as edge weights
2. Find $(k-1)$-nearest neighbor graph
3. Construct the second power of NNG
4. Find a maximal independent set (seeds)
5. Form blocks with the seeds and their neighbors in NNG
6. Assign remaining units to a block containing any neighbor



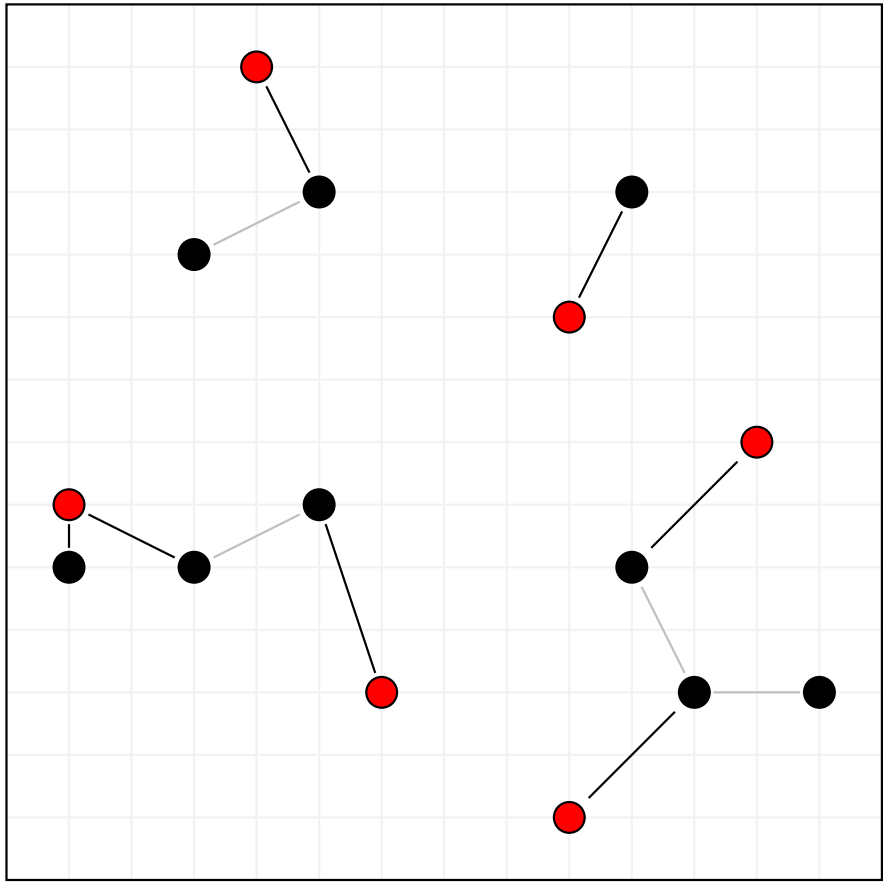
The AppOpt algorithm

Input:

- Units' covariates
- Distance metric
- Minimum block size: $k = 2$

Procedure:

1. A undirected complete graph with distances as edge weights
2. Find $(k-1)$-nearest neighbor graph
3. Construct the second power of NNG
4. Find a maximal independent set (seeds)
5. Form blocks with the seeds and their neighbors in NNG
6. Assign remaining units to a block containing any neighbor



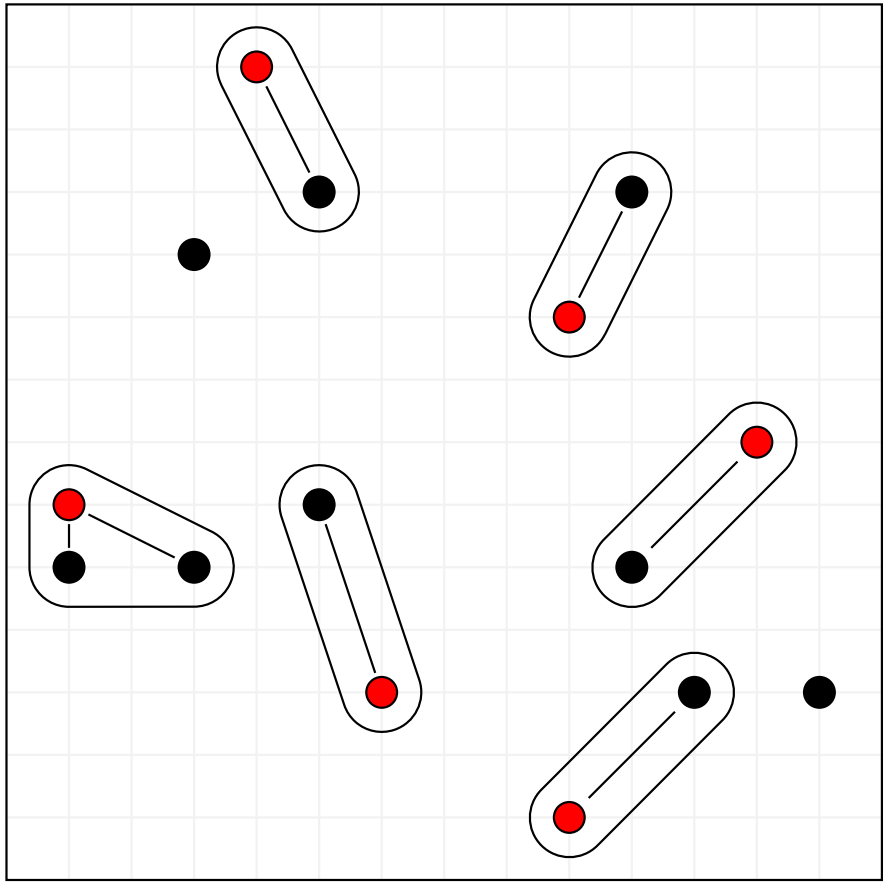
The AppOpt algorithm

Input:

- Units' covariates
- Distance metric
- Minimum block size: $k = 2$

Procedure:

1. A undirected complete graph with distances as edge weights
2. Find $(k-1)$-nearest neighbor graph
3. Construct the second power of NNG
4. Find a maximal independent set (seeds)
5. Form blocks with the seeds and their neighbors in NNG
6. Assign remaining units to a block containing any neighbor



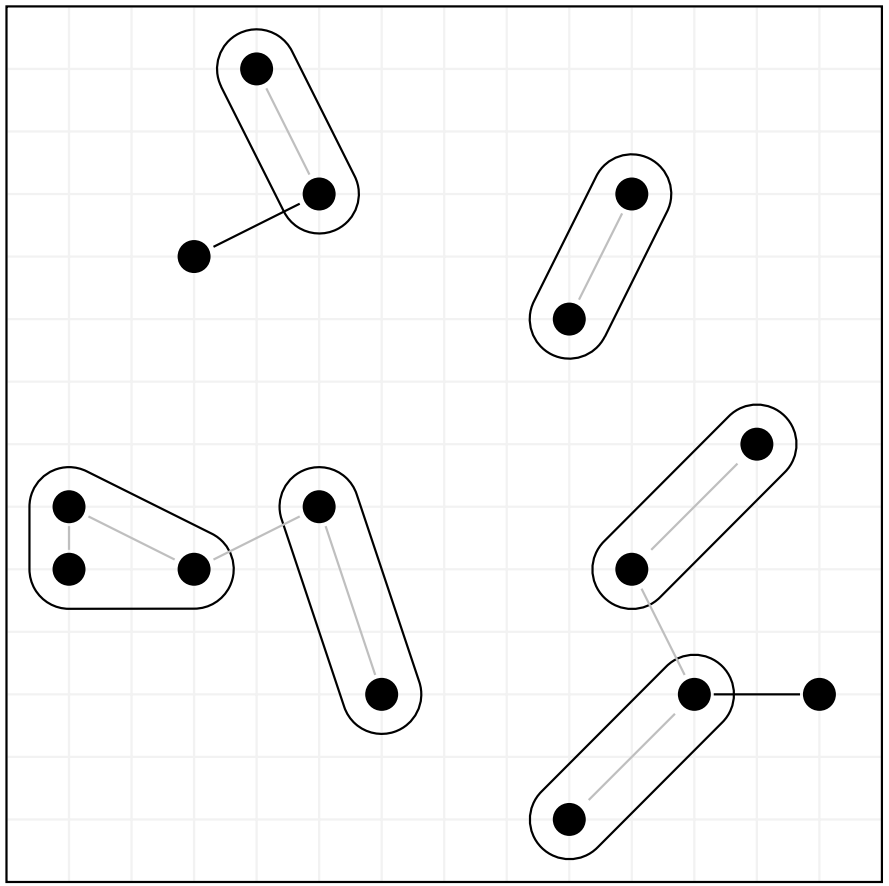
The AppOpt algorithm

Input:

- Units' covariates
- Distance metric
- Minimum block size: $k = 2$

Procedure:

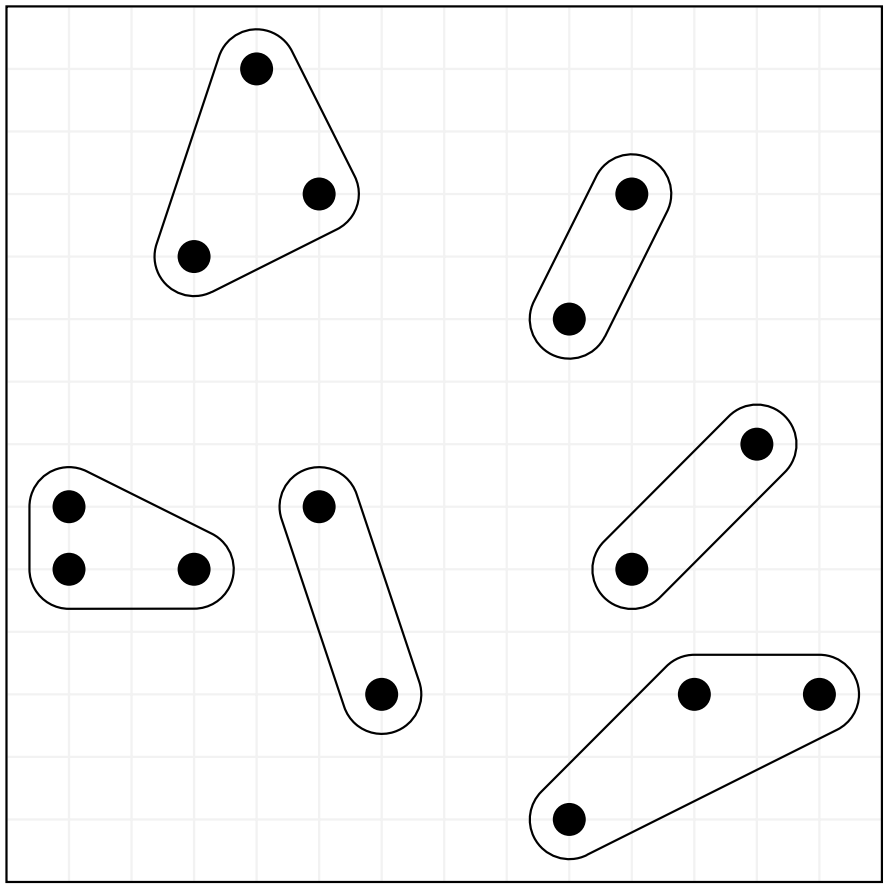
1. A undirected complete graph with distances as edge weights
2. Find $(k-1)$-nearest neighbor graph
3. Construct the second power of NNG
4. Find a maximal independent set (seeds)
5. Form blocks with the seeds and their neighbors in NNG
6. Assign remaining units to a block containing any neighbor



The AppOpt algorithm

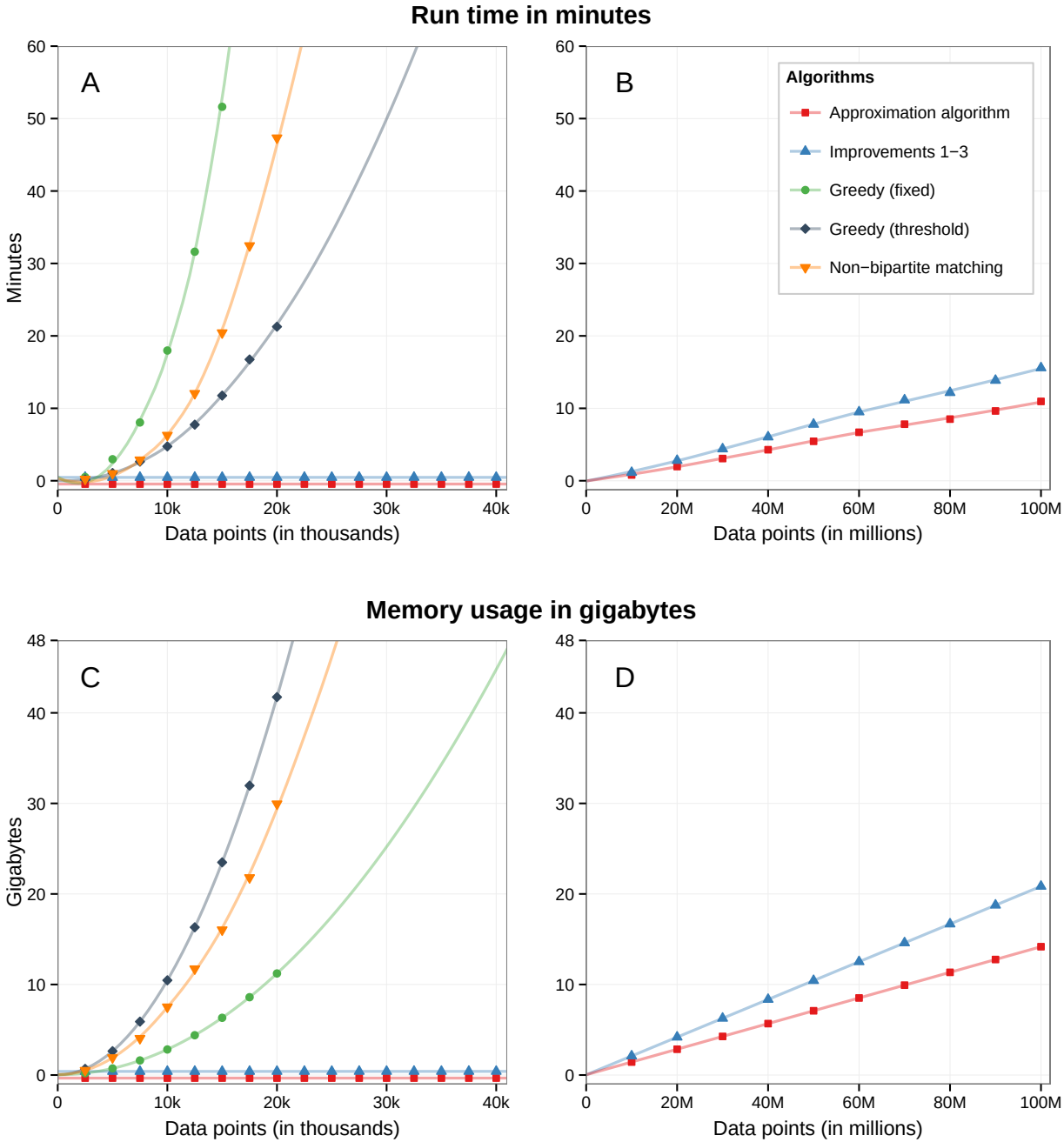
Input:

- Units' covariates
- Distance metric
- Minimum block size: $k = 2$



Procedure:

1. A undirected complete graph with distances as edge weights
2. Find $(k-1)$-nearest neighbor graph
3. Construct the second power of NNG
4. Find a maximal independent set (seeds)
5. Form blocks with the seeds and their neighbors in NNG
6. Assign remaining units to a block containing any neighbor


Run time in minutes
A
B
Minutes
Data points (in thousands)
Data points (in millions)
Algorithms
Approximation algorithm
Improvements 1−3
Greedy (fixed)
Greedy (threshold)
Non−bipartite matching
Memory usage in gigabytes
C
D
Gigabytes
Data points (in thousands)
Data points (in millions)

Properties

- Unless $P = NP$, no polynomial-time $(2 - \epsilon)$ approximation algorithm exists for any $\epsilon > 0$
- Validity: the blocking algorithm produces a threshold blocking: $\mathbf{b}_{alg} \in \mathbf{B}_k$
- Complexity: the blocking algorithm terminates in polynomial time using $O(kn)$ space
- Approximate optimality: blocking algorithm is a 4-approximation algorithm:

$$\max_{ij \in E(\mathbf{b}_{alg})} c_{ij} \leq 4\lambda.$$

- Local approximate optimality: Let $\mathbf{b}_{sub} \subseteq \mathbf{b}_{alg}$ be any subset of blocks from a blocking constructed by the algorithm. Define $V_{sub} = \bigcup_{V_x \in \mathbf{b}_{sub}} V_x$ as the set of all vertices contained in the blocks of $\mathbf{b}_{sub}$. Let λ_{sub} denote the maximum edge cost in an optimal blocking of V_{sub}. The subset of blocks is an approximately optimal blocking of V_{sub}:

$$\max_{ij \in E(\mathbf{b}_{sub})} c_{ij} \leq 4\lambda_{sub}.$$

Summary

- Fast algorithm:
  - NNG plus $O(d^0 kn)$ time and $O(d^0 kn)$ space
  - K-d trees NN: $O(2^d kn \log n)$ expected time, $O(2^d kn^2)$ worst time, and $O(kn)$ storage
  - Compare with bipartite, network flow methods:
    - e.g., Derigs: $O(n^3 \log n + dn^2)$ worst time and $O(d^0 n^2)$ space
- Closer to clustering than traditional blocking methods
- Important for separating design from analysis
- Lots of questions about best way to handle estimation
  - Design based estimators: Difference of means; Horvitz-Thompson estimator; double Hájek estimator
  - Probably do want to run a model on the blocked data. What if there is heterogeneity by blocks? $\frac{p}{n} \neq 0$

Joint Estimation Method

Borrow strength from the observational data, but:

- If the observational data is not useful, it should be weighted little. Estimates should be based on the RCT
- If the observational data contains useful information, it will be positively weighted
- One can estimate either sample or population parameters

Estimation Method

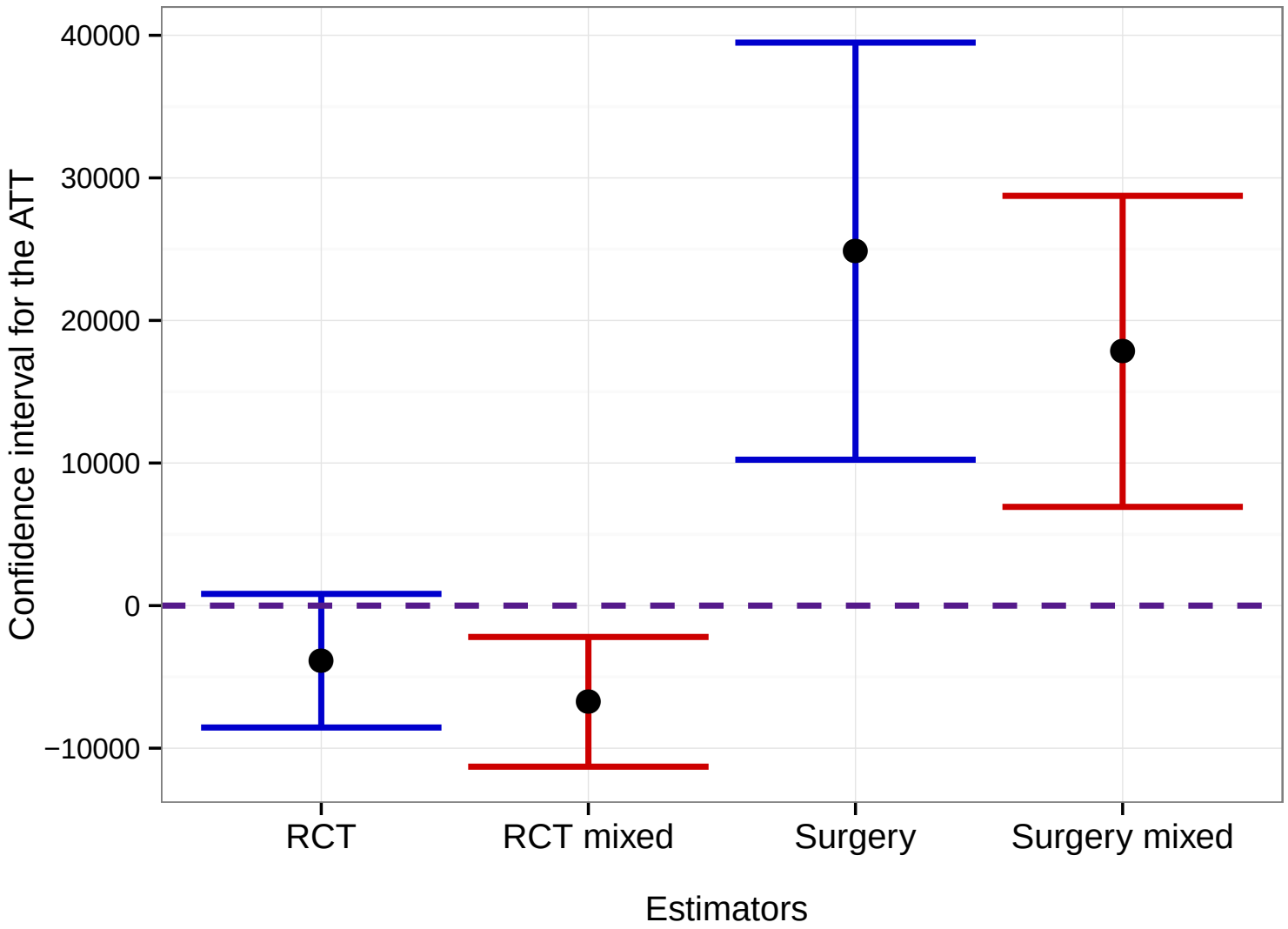
$$\left(\lambda \cdot \beta_{RCT}^{training} + (1-\lambda) \cdot \beta_{NRS} - \beta_{RCT}^{test}\right)^2$$

$$\Rightarrow \lambda = \underset{\lambda \in \{\lambda:\, 0 \leq \lambda \leq 1\}}{\text{argmin}} \left\{ \left(\lambda \cdot \beta_{RCT}^{training} + (1-\lambda) \cdot \beta_{NRS} - \beta_{RCT}^{test}\right)^2 \right\}$$

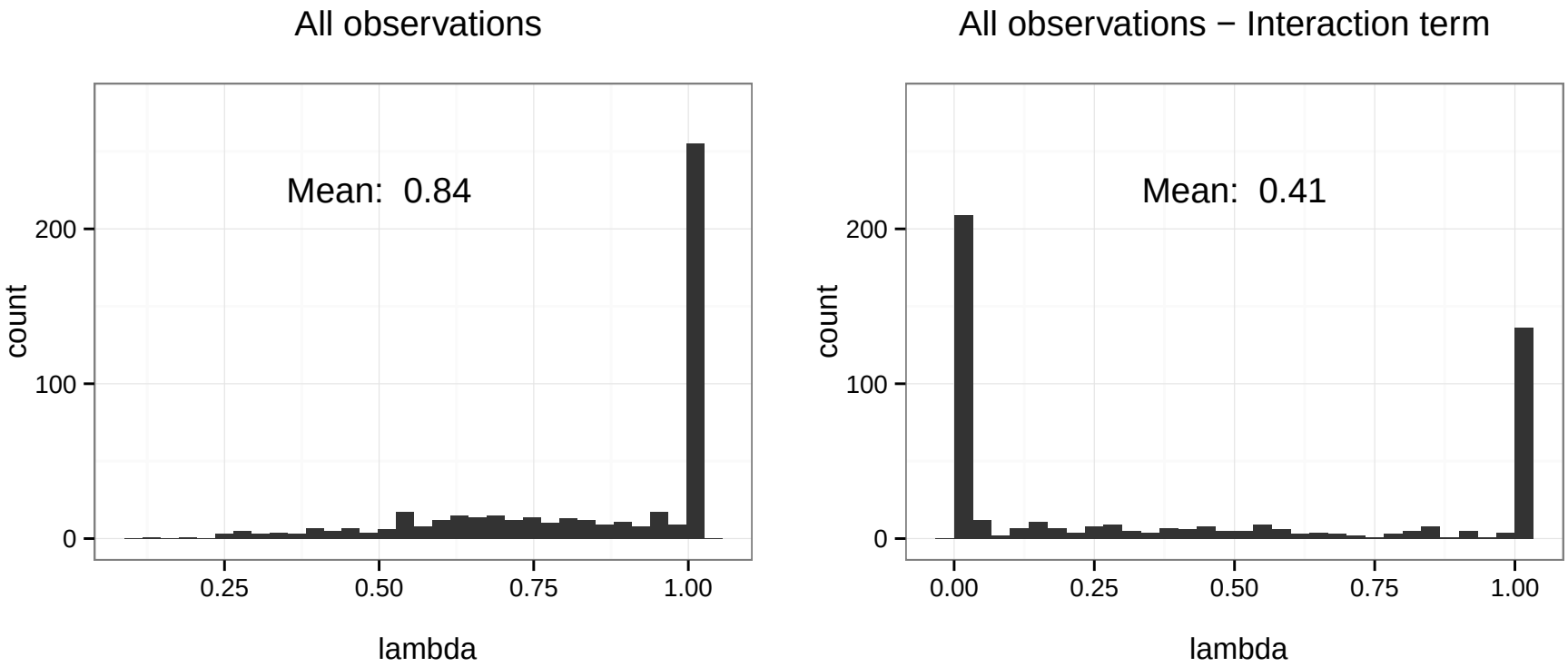
Where, $0 \leq \lambda \leq 1$
The optimal weight to assign $\beta_{RCT}^{training}$ is,

$$\hat{\lambda} = \min\left(1, \max\left(0, \frac{\beta_{RCT}^{test} - \beta_{NRS}}{\beta_{RCT}^{training} - \beta_{NRS}}\right)\right)$$

Combined Estimates



The distribution of λ



The distribution of the weight given to the RCT estimate

Bibliography I

David A. Freedman. On regression adjustments in experiments with several treatments. The annals of applied statistics, 2(1):176–196, 2008.

Robert Greevy, Bo Lu, Jeffrey H. Silber, and Paul Rosenbaum. Optimal multivariate matching before randomization. Biostatistics, 5(4):263—-275, 2004.

Guido W. Imbens. Experimental design for unit and cluster randomized trials. Working Paper, 2011.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. Annals of Applied Statistics, 2012.

Luke W. Miratrix, Jasjeet S. Sekhon, and Bin Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments. Journal of the Royal Statistical Society, Series B, 75(2):369–396, 2013.

Ryan T Moore. Multivariate continuous blocking to improve political science experiments. Political Analysis, 20(4):460–479, 2012.

Kari Lock Morgan and Donald B Rubin. Rerandomization to improve covariate balance in experiments. Annals of Statistics, 40(2):1263–1282, 2012.