500 Class 06

https://thomaselove.github.io/500-2024/

2024-02-22

What's in these Slides?

- Designing Observational Studies (Rubin, 2001)
- Discussion of Project Proposals
- Kubo et al. (2020) as an example for OSIA
- Some Extensions to Propensity Matching (originally in Slides Set 5)

Section 1

Designing Observational Studies (Rubin 2001)

On Designing Observational Studies

- Exert as much experimental control as possible
- Carefully consider the selection process
- Actively collect data to reveal potential biases

  "Care in design and implementation will be rewarded with useful and clear study conclusions... Elaborate analytical methods will not salvage poor design or implementation of a study." – NAS report (quoted in Rosenbaum p. 368)

But **HOW**?

Rubin 2001

On Designing an Observational Study with the Propensity Score

Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation

DONALD B. RUBIN

Rubin 2001 Abstract

Abstract. Propensity score methodology can be used to help design observational studies in a way analogous to the way randomized experiments are designed: without seeing any answers involving outcome variables. The typical models used to analyze observational data (e.g., least squares regressions, difference of difference methods) involve outcomes, and so cannot be used for design in this sense. Because the propensity score is a function only of covariates, not outcomes, repeated analyses attempting to balance covariate distributions across treatment groups do not bias estimates of the treatment effect on outcome variables. This theme will the primary focus of this article: how to use the techniques of matching, subclassification and/or weighting to help design observational studies. The article also proposes a new diagnostic table to aid in this endeavor, which is especially useful when there are many covariates under consideration. The conclusion of the initial design phase may be that the treatment and control groups are too far apart to produce reliable effect estimates without heroic modeling assumptions. In such cases, it may be wisest to abandon the intended observational study, and search for a more acceptable data set where such heroic modeling assumptions are not necessary. The ideas and techniques will be illustrated using the initial design of an observational study for use in the tobacco litigation based on the NMES data set.

Keywords: balance, matching, subclassification

Designing an Observational Study without access to the outcome data

- Propensity score methods can be used to help design the OS without seeing any outcomes.
  - Propensity score is a function only of covariates, not of outcomes.

The key insight from Rubin (2001)

Repeated analyses attempting to balance covariate distributions across treatment groups **do not bias** estimates of the treatment's effect on outcome variables.

Designing Observational Studies

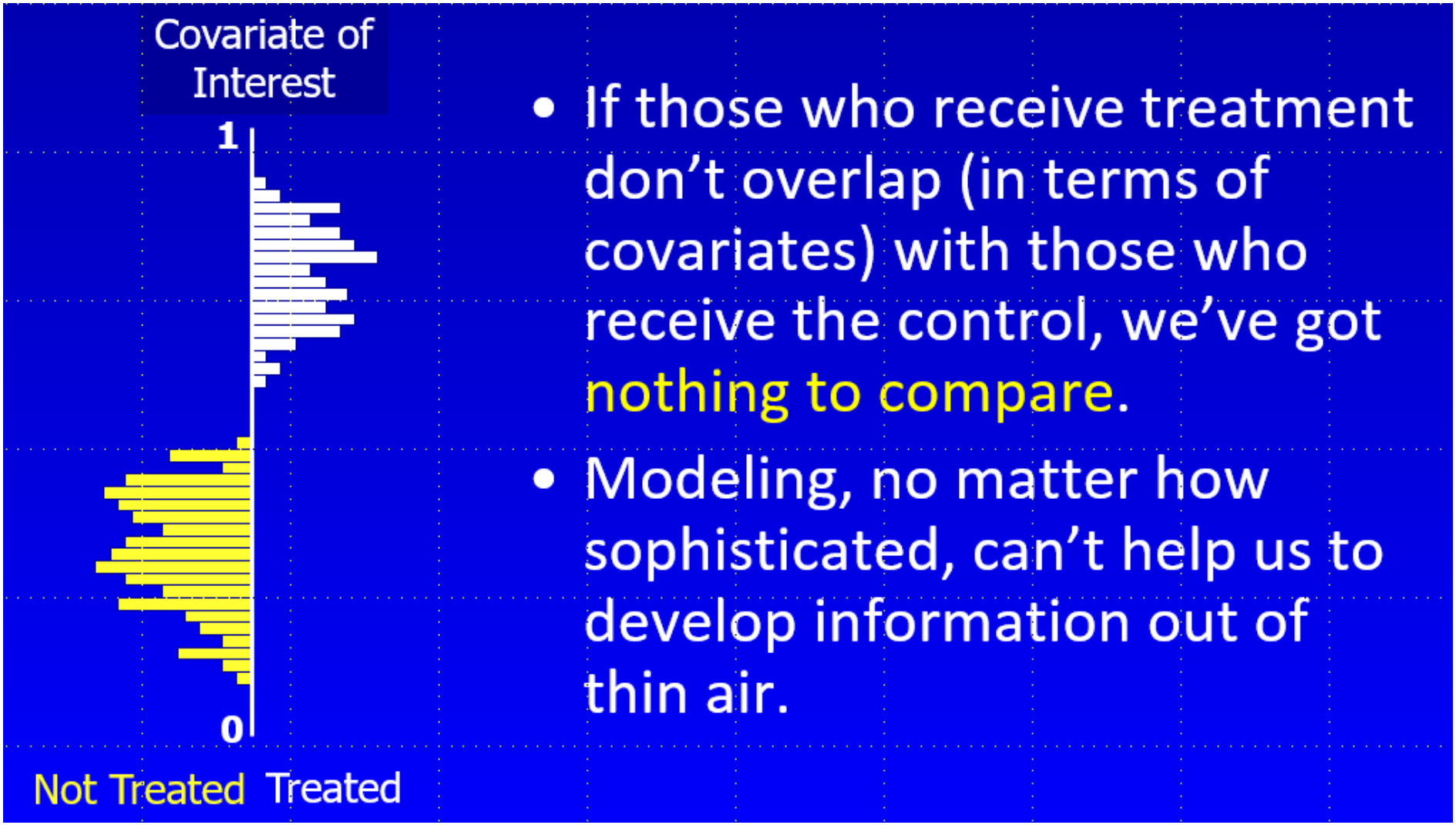
- The Importance of Covariate (PS) Overlap
- How To Check for Overlap Effectively
- Designing Like You're Doing an Experiment
- Using Matching, Subclassification and Weighting
- Propensity Scores are "Fair Game" - No Outcomes!

In order to extract information on treatment effect from an observational study, we need to be able to compare "identical" people who receive different treatments.

Goal: Use propensity scores to assemble treatment groups that have comparable distributions on all measured covariates.

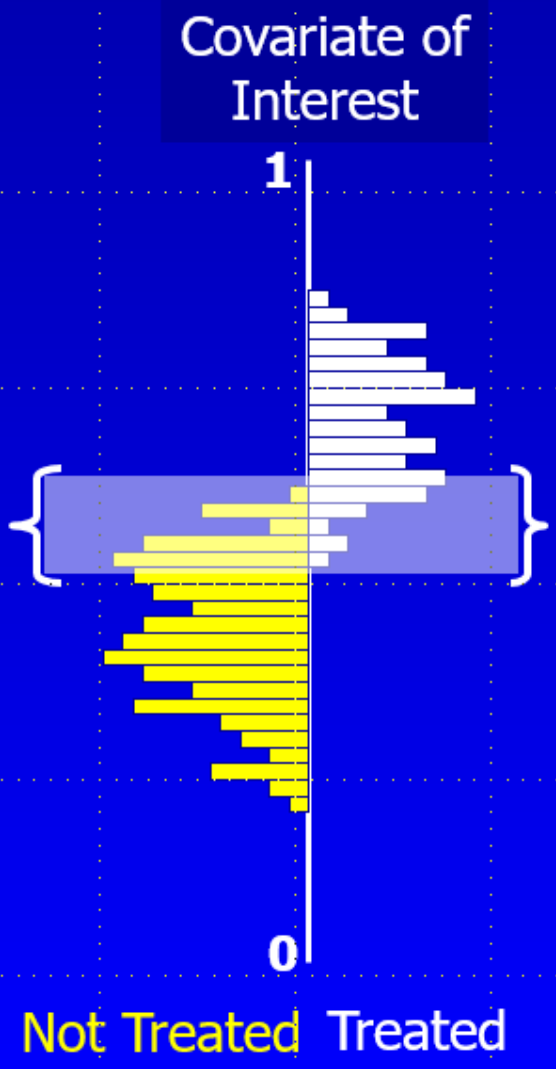
Issue 1: Overlap

How much overlap in the covariates do we want?



- If those who receive treatment don't overlap (in terms of covariates) with those who receive the control, we've got nothing to compare.
- Modeling, no matter how sophisticated, can't help us to develop information out of thin air.

What if the exposure groups overlap, but minimally?



- No help.
- The information available to infer treatment effect will reside almost entirely in the few patients who overlap.
- Need to think hard about whether useful inferences will be possible.

Initial Phases of Ideal Study Design

Specify population, exposures/treatments, outcomes and covariates.

- Collect treatment and covariate information, and model treatment assignment with the propensity score.
- Use propensity scores (through matching, stratification, reweighting) to reduce bias.
- Check for covariate balance across the treatment groups and iterate through process.
  - If the treatment and control groups have the same distribution of propensity scores, then they have the same distribution of all observed covariates, just as in a randomized experiment.
  - Of course, propensity scores are only guaranteed to balance the observed covariates, while randomized experiments can stochastically balance unobserved, as well.

Rich and Poor Covariate Sets

- With a rich set of covariates, adjustments for hidden covariates may be less critical.
- With less rich covariate sets, we may need to do more, say, try to find an instrument.

As Rubin mentions in the Abstract, our conclusion after the initial design stage may be that the treatment and control groups are too far apart to produce reliable effect estimates without heroic modeling assumptions.

Techniques for Initial Observational Study Design using Propensity Scores

- Matching
- Subclassification / Stratification
- Weighting

Goal: Assemble groups of treated and control units such that within each group the distribution of covariates is balanced.

Allows us to attribute outcome differences to the effect of treatment vs. control.

Why Work this Hard in the Initial Design Stage?

- Options narrow as an investigation proceeds.
- No harm, no foul.
  - Since no outcome data are available to the PS, nothing based on the PS here biases estimation of treatment effects.
- Balancing covariates / PS makes subsequent model-based adjustments more reliable.

Key point is that model adjustments can be extremely unreliable when the treatment groups are far apart on covariates. So we need to avoid that.

"Balancing" helps in terms of assessing covariance, relative risk, subsequent adjustments, etc.

Propensity Score **Matching** in the Design of an Observational Study

- Pair up treated and control subjects with similar values of the propensity score, discarding all unmatched units.
  - Not limited to 1-1 matches, can do 1-many, etc.
- Can find an *optimal* full match using `optmatch` in R, without discarding any units, then follow with adjustments.
  - Technically more valid, but difficult sell in practice.
- Common: One-one Mahalanobis matching within calipers defined by logit(propensity).

Propensity Score **Subclassification** in the Design of an Observational Study

- Rank all subjects by their propensity score and then create subclasses by imposing boundaries.
- Subclasses therefore have treated and control units with similar values of propensity score.
- Often use 5 subclasses of equal size should remove 90% or more of the bias due to the observed covariates in the propensity score.

Propensity Score **Weighting** in the Design of an Observational Study

Estimate propensity scores for each subject, so that PS = prob(treatment received | covariates)

Rubin describes the ATE approach to weighting...

- Weights for treated subjects: $\frac{1}{PS}$.
- Weights for control subjects: $\frac{1}{1-PS}$

When Can We Move On?

Three conditions which must all apply for regression adjustment to be trustworthy:

1. Difference in the means of linear propensity score [logit(PS)] in the two groups being compared must be small.
2. Ratio of variances of linear propensity scores in the two groups must be close to 1.
3. Ratio of variances of the "residuals" of the covariates after PS adjustment close to 1.

These are what I have referred to as "Rubin's Rules"...

Three Rules (page 174, Rubin 2001)

In particular, there are three basic distributional conditions that in general practice must simultaneously obtain for regression adjustment (whether by ordinary linear regression, linear logistic regression, or linear-log regression) to be trustworthy. If any of these conditions is not satisfied, the differences between the distributions of covariates in the two groups must be regarded as substantial, and regression adjustment will be unreliable and cannot be trusted. These conditions are:

1. The difference in the means of the propensity scores in the two groups being compared must be small (e.g., the means must be less than half a standard deviation apart), unless the situation is benign in the sense that: (a) the distributions of the covariates in both groups are nearly symmetric, (b) the distributions of the covariates in both groups have nearly the same variances, and (c) the sample sizes are approximately the same.
2. The ratio of the variances of the propensity score in the two groups must be close to one (e.g., 1/2 or 2 are far too extreme).
3. The ratio of the variances of the residuals of the covariates after adjusting for the propensity score must be close to one (e.g., 1/2 or 2 are far too extreme); "residuals" precisely defined shortly.

Assessing Balance on the *Linear* rather than *Raw* Propensity Score

- *logit*(PS) is more relevant for assessing whether linear modeling adjustments work.
- *logit*(PS) tend to have more benign (variances closer, greater symmetry) distributions.
- *logit*(PS) are more closely related to benchmarks in the literature on adjustments for covariates based on linearity assumptions.

Putting Rubin's Rule 1 into operation

1. Difference in the means of the propensity scores in the two groups being compared.
    - Estimate propensity scores for all subjects.
    - Take logit(PS) for each subject (normalize).
    - Find SD = standard deviation of logit(PS) across all subjects (treated and control).
    - Mean logit(PS) for treated group should be within 0.5 SD of control group's mean logit(PS).
    - Often we calculate a standardized difference here.

```
rubin1.unadj <- with(toy, abs(100*(mean(linps[treated==1]) -
                                   mean(linps[treated==0]))
                                     sd(linps)))
```

Putting Rubin's Rule 2 into operation

2. Variance ratio of propensity scores in the two groups being compared should be close to 1.
    - Estimate propensity scores for all subjects.
    - Take logit(PS) for each subject (normalize).
    - Find variance of logit(PS) across treated subjects, and divide it by the variance of logit(PS) across control subjects.
    - Variance ratio should be close to 1. Ratios of 0.5 and 2.0 are far too extreme: we often try for (4/5, 5/4).

```
rubin2.unadj <-with(toy, var(linps[treated==1]) /
                         var(linps[treated==0]))
```

Putting Rubin's Rule 3 into operation

3. Variance ratio of "residuals" close to 1.
    - Estimate propensity scores for all subjects.
    - For each covariate, regress the original value of the covariate for each subject on logit(PS) and take the residual of this regression.
    - For each covariate, divide variance of the residuals within treatment group by variance of the residuals within control group.
    - For each covariate, this variance ratio should also be close to 1 (2 or 0.5 are, again, far too extreme).

rubin3 function built for the toy example

```
## General function rubin3 to help calculate Rubin's Rule 3
rubin3 <- function(data, covlist, linps) {
  covlist2 <- as.matrix(covlist)
  res <- NA
  for(i in 1:ncol(covlist2)) {
    cov <- as.numeric(covlist2[,i])
    num <- var(resid(lm(cov ~ data$linps))[data$treated==1])
    den <- var(resid(lm(cov ~ data$linps))[data$treated==0])
    res[i] <- round(num/den, 3)
  }
  names(res) <- names(covlist)
  print(res)
}
```

National Medical Examination Survey

- Large nationally representative data base of nearly 30,000 adults, calendar year 1987
- Modern related efforts are folded into NHANES

Goal: objective observational study on the causal effects of smoking and the effect of the tobacco companies' alleged misconduct

NMES Covariates for Smoking Study

- Age, Sex, Race, Marital Status, Education, etc.
- Detailed smoking information
  - Classification of subjects as never smokers, former smokers and current smokers
  - Further classifications possible by length and density of smoking behaviors
  - Also can look at years since quitting for former smokers

NMES Objects of Inference

- Smoking Attributable Fractions
- Conduct Attributable Fractions
- Relative Expenditure Risks

All based on comparisons of specific health-related expenditures (or disease rates)

Comparisons of smokers with "never smokers" who have same covariate values, as a function of dosage and covariates

Rubin's Main Example

Design Goal: Create samples of smokers and never smokers in NMES with the same multivariate distribution of covariates.

- Males and Females treated separately.
- We'll focus first on Male "Current Smokers" vs. Male "Never Smokers"
  - 3510 Male "Current Smokers" in the pool
  - 4297 Male "Never Smokers"" as controls
- Fit propensity for "current smoker" to these people, via logistic regression with sampling weights

Separate models were built for "former vs. never (Males)" and the two analogous comparisons of Females.

Propensity Model: 146 Covariates

Variables Used in Propensity Model	Description
Seatbelt	5 levels of reported seat belt use
Arthritis	Whether reported suffering from arthritis
Census Division	9 census regions
Champ Insurance	Whether have military insurance
Diabetes	Doctor ever told having diabetes
Down time	6 levels of reported emotional down time
Dump time	6 levels of reported in the dumps time
Employment	Indicating employment status each quarter
English	English is a primary language
Retirement	Indicator for retirement status
Number of Friends	7 levels measuring the number of friends
Membership in Clubs	6 levels measuring memberships in clubs
Education	Completed years of education
HMO coverage	Indicating HMO coverage each quarter

Propensity Model: 146 Covariates

High blood pressure	Doctor ever told having high blood pressure
Industry Code	14 Industry codes
Age	Age of the respondent
Labor Union	Indicator for a member of labor union
Log Height	Natural Logarithm of height
Log Weight	Natural Logarithm of weight
Marital Status	Marital status in each quarter
Medicaid	On medicaid (each quarter)
Medicare	On medicare (each quarter)
Occupation	Occupation code (13 levels)
Public Assistance	Other public assistance program (each quarter)
Friends over	Frequency of having friends over (7 levels)
Physical Activity	Indicator variable for physically active
Population density	3 levels
Poverty Status	6 levels
Pregnant 1987	Pregnancy status in 1987 (women)
Private Insurance	Other private insurance (each quarter)
Race	4 levels

Propensity Model: 146 Covariates

Race	4 levels
Rated Health	5-point self rating of health status
Home ownership	Indicator for owning home
Rheumatism	Indicator for suffering from rheumatism
Share Life	Indicator variable for having somebody to share their life
Region	4 levels of region of the country
MSA	4 levels indicating types of metropolitan statistical area
Risk	General risk taking attitude (5 levels)
Uninsured	Indicator for lack insurance (each quarter)
Veteran	Indicator for veteran status
Incapler	Survey weight in NMES database
Agesq	Age*Age
Educat.sq	Education*Education
Age_wt	Age*Logwt
Age_educt	Age*Education
Age_ht	Age*Loght
Educat_wt	Education*Logwt

Propensity Model: 146 Covariates

Variables Used in Propensity Model	Description
Educat_ht	Education*Loght
Loght_logwt	Loght*Logwt
Loghtsq	Loght*Loght
Logwtsq	Logwt*Logwt

Assessing Overlap Step 1: Looking for Mean Bias

Bias B = standardized difference in the means of logit(propensity scores) between current smokers and never smokers for males

- We want the bias in the propensity score to be small, no greater than 0.50 in absolute value.
- Here, mean propensity score Bias B = 1.09 (109%)
- In fact standardized difference > 0.5 (50%) for many of the individual covariates, as well.

Assessing Overlap Step 2: Comparing Variances

Ratio R = ratio of the variances of logit(propensity scores) between current smokers and never smokers for males.

- We want the variances to be homogeneous, so the ratio should be close to 1 (1/2 and 2 are far too extreme).
- Here, variance ratio for logit(PS) is R = 1.00
- Could look at ratio of individual covariate variances, also. (In fact, `MatchBalance` does this.)

Assessing Overlap Step 3: Comparing Residuals

Regress each covariate on logit(PS) and look at ratio of variances of residuals for current smokers to variance of residuals for never smokers within the male population.

- Here, we get a separate result for each of the 146 covariates. We want results near 1.00
  - 57% of the covariates had their residual ratio between 4/5 and 5/4
  - 5% of covariates had their residual ratio below 1/2, or above 2

Excerpt from Rubin's Table 2 (page 179)

Table 2. Estimated propensity scores on the logit scale for "smokers" versus never smokers in full NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Treated Group	B	R	$\leq 1/2$	$>1/2$ and $\leq 4/5$	$>4/5$ and $\leq 5/4$	$>5/4$ and ≤ 2	>2
Male Current $N=3{,}510$	1.09	1.00	3	9	57	26	5

B = Bias, R = Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.

Interpretation

> *"... [A]ny linear (or part linear) regression model cannot be said to adjust reliably for these covariates, even if they were perfectly normally distributed. ... B [is] greater than 1/2, and many of the value of R for the residuals of the covariates are outside the range (4/5, 5/4)."*

Similar Results for the Other Study Comparisons

Table 2. Estimated propensity scores on the logit scale for "smokers" versus never smokers in full NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Treated Group	B	R	$\leq 1/2$	$>1/2$ and $\leq 4/5$	$>4/5$ and $\leq 5/4$	$>5/4$ and ≤ 2	>2
Male Current $N=3{,}510$	1.09	1.00	3	9	57	26	5
Male Former $N=3{,}384$	1.06	0.82	2	15	61	15	7
Female Current $N=3{,}434$	1.03	0.85	1	15	59	23	2
Female Former $N=2{,}657$	0.65	1.02	5	7	85	7	5

$B=$ Bias, $R=$ Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.

All four comparisons indicate the need for propensity score adjustments.

Mahalanobis Matching within PS Calipers

For the 3510 male current smokers, 3510 "matching" male never smokers were chosen from the pool of 4297 male never smokers.

- Method: Mahalanobis metric matching within propensity score calipers ($\pm$ 0.2 of the standard deviation of linear propensity scores)
  - Mahalanobis distance variables were: age, education, body mass index, and sampling weight.
  - Some of these are survey results, mostly (but not completely) in categories.
- In this case, there were no current smoker Males that could not be matched within the PS calipers to never smoker Males.
  - What if there had been a treated subject whose propensity score was not "matchable"?
  - What if the "donor pool" of never smokers had been empty for one of the current smokers?

Impact of Matching on Overlap

Male Current Smokers vs. Male Never Smokers

Scenario	Bias, B	Variance Ratio, R
Before Matching	1.09	1.00
After Matching	0.08	1.16

Residual Variance Ratios (% in range)

Range	Before Match	After Match
≤ 0.5	3	1
$(\frac{1}{2}, \frac{4}{5}]$	9	3
$(\frac{4}{5}, \frac{5}{4}]$	**57**	**90**
$(\frac{5}{4}, 2]$	26	6
> 2	5	0

Matching's Impact on Overlap

Male Former Smokers vs. Male Never Smokers

Scenario	B	R	Res. VR in $(\frac{4}{5}, \frac{5}{4}]$
Before Match	1.06	0.82	61% of covariates
After Match	0.04	0.99	94%

Female Comparisons re: Matching

Female **Current** Smokers vs. Female Never Smokers

Scenario	B	R	Res. VR in $(\frac{4}{5}, \frac{5}{4}]$
Before Match	1.03	0.85	59% of covariates
After Match	0.04	0.94	93%

Female **Former** Smokers vs. Female Never Smokers

Scenario	B	R	Res. VR in $(\frac{4}{5}, \frac{5}{4}]$
Before Match	0.65	1.02	85%
After Match	0.06	1.02	91%

Re-estimating PS using Matched Subjects Only

Original propensity score estimate used all of the subjects, including those subjects who wound up being unused controls, once we matched.

- Here, they are no longer concerned with unmatched "never smokers" so they re-estimate the propensity score using only the matched samples, then look at the remaining covariate imbalance.

Group	B	R	Res. VR in $(\frac{4}{5}, \frac{5}{4}]$
Male, Current	0.39	1.33	88%
Male, Former	0.32	1.33	95%
Female, Current	0.35	1.18	92%
Female, Former	0.31	1.09	91%

Looks better. Suppose we are still not satisfied, though.

Subclassification of Matched Samples

Suppose we are still not satisfied...

Create two equal-size (weighted) subclasses, low and high on the linear PS.

- Treated and Control subjects with low PS are to be compared to each other.
- Treated and Control subjects with high PS are to be compared to each other.
- Weighted average of two comparisons yields the result.

Subclassification as Re-Weighting

- For the treated subjects, the new weights implied by this subclassification are the total (weighted) number of treated and controls in that subclass, divided by the total (weighted) number of treated subjects.
- For the control subjects, weights are the subclass total of treated & controls divided by subclass controls.

Leads to a weighted PS analysis that reflects the additional balance due to subclassification.

- The same idea for weighting works no matter how many subclasses
  - One subclass is what we've had - no subclassification adjustment, just matching.
  - We'll also look at the impact of incorporating 2, 4, 6, 8, or 10 subclasses after matching...

Current vs. Never Smoking Males: Overlap

Matching + Post-Matching Subclassification

Subclasses	B	R	Res. VR in $(\frac{4}{5}, \frac{5}{4}]$
1	0.39	1.33	88%
2	0.18	1.36	98%
4	0.10	1.25	99%
6	0.09	1.30	100%
8	0.08	1.16	100%
10	0.07	1.12	100%

How Far Can We Go?

We can obtain dramatic reduction in initial bias through this sort of subclassification, and we can carefully pick out just how many subclasses will be most helpful in getting the job done.

We can even do Weighted Propensity Score Analysis (using infinitely many subclasses)

- Form ATE weights directly from the estimated propensity score without subclassification.
  - Weight for treated subject: inverse of his/her propensity score (times his/her NMES weight)
  - Weight for control subject: inverse of 1 minus his/her propensity score (times NMES weight)
  - Caveat: Can get unrealistically extreme weights when estimated PS is near zero or one.

Current vs. Never Smoking Males: Overlap

Analysis	B	R	Res. VR in $(\frac{4}{5}, \frac{5}{4}]$
Full Sample	1.09	1.00	57%
Match full PS	0.08	1.16	90%
Match new PS	0.39	1.33	88%
Match, then 2 subclasses	0.18	1.36	98%
4 subclasses	0.10	1.25	99%
6 subclasses	0.09	1.30	100%
8 subclasses	0.08	1.16	100%
10 subclasses	0.07	1.12	100%
Match, then Weight	0.03	1.19	100%

Why Work this Hard?

- If substantial balance in covariates is obtained in this initial design stage, the exact form of the modeling adjustment is not critical.
- Similar treated and control covariate distributions implies only limited model-based sensitivity.

Why doesn't this introduce a bias for our eventual conclusions and analytic results?

Why can we get away with this?

- We're not affecting our conclusions in a biased way, because we don't look at outcomes here.

Why can we get away with this?

- We're not affecting our conclusions in a biased way, because we don't look at outcomes here.
- In fact, I've yet to specify the outcomes.

NMES Outcomes for Smoking Study

- Health-care expenditures of various types
- Occurrence of various smoking-related diseases

Remember, these outcomes are never seen during the design process.

Section 2

Discussion of Project Proposals

Project 1 Proposals

- Population (How as the sample selected?)
- Outcome (or response)
- Treatment (or exposure)
- Covariates

Section 3

Kubo et al. (2020) as an example for OSIA

Kubo et al. (2020) Paper

Effects of preoperative low-intensity training with slow movement on early quadriceps weakness after total knee arthroplasty in patients with knee osteoarthritis: a retrospective propensity score-matched study

- **Population**
- **Outcome**
- **Treatment**
- **Covariates**

Key Words and Abbreviations

Key Words (from Abstract): Exercise preconditioning, Ischemic preconditioning, Ischemia-reperfusion injury, Knee swelling, Low-intensity training, Prehabilitation, Quadriceps weakness, Slow movement, Thigh swelling, Total knee arthroplasty

Abbreviations

TKA: Total knee arthroplasty; QW: Quadriceps weakness; IR: Ischemiareperfusion; IPC: Ischemic preconditioning; EPC: Exercise preconditioning; LST: Low-intensity resistance exercise with slow movement and tonic force generation; QST: Quadriceps strength test; TUG: Timed up and go test; SCT: Stair climb test; JKOM: Japanese Knee Osteoarthritis Measure; VAS: Visual analog scale; SMD: Standardized mean difference

Background

Background: Severe and early quadriceps weakness (QW) after total knee arthroplasty (TKA), which is caused by acute inflammation resulting from surgical trauma and tourniquet-induced ischemia-reperfusion (IR) injury, can be especially problematic. We focused on tourniquet-induced IR injury, because it has been shown to be preventable through ischemic and exercise preconditioning. Low-intensity resistance exercise with slow movement and tonic force generation (LST) share some similarities with ischemic and exercise preconditioning. The present study primarily aimed to clarify the efficacy of preoperative LST program as prehabilitation for early QW among patients with TKA using propensity score matching analysis.

Methods

Methods: This single-center retrospective observational study used data from patients with knee osteoarthritis (n = 277) who were scheduled to undergo unilateral TKA between August 2015 and January 2017. Those with missing outcome data due to their inability to perform tests were excluded. The LST group included participants who performed LST and aerobic exercise (LST session) more than seven times for three months prior to surgery. The control group included participants who performed less than eight LST sessions, a general and light exercise or had no exercise for three months prior to surgery. Knee circumference, thigh volume, knee pain during quadriceps strength test (QST) and timed up and go test (TUG), quadriceps strength, and TUG were measured before and 4 days after surgery. Knee swelling, thigh swelling, Δknee pain, QW, and ΔTUG were determined by comparing pre- and postoperative measurements.

Check 1

Can we describe the ...

- Population
- Outcome
- Treatment
- Covariates

Statistical Analysis Section (Start)

Statistical analysis was conducted using the IBM SPSS version 26 statistical software package (IBM Corp., Armonk, N.Y., USA). Participants were divided into the LST group and control group. The LST group included participants who performed category 1 sessions (LST and aerobic exercise) more than seven times for three months prior to surgery. The control group included participants who performed less than eight category 1 sessions, category 2 sessions (a general and light exercise) or had no prehabilitation (no exercise) for three months prior to surgery. Propensity score matching was used to balance group characteristics that could affect the LST program's instructions and formulae. Propensity scores were estimated using a logistic regression model where treatment status was regressed on age, gender, body mass index, and preoperative measurements, including quadriceps strength of the affected leg, knee pain during the QST and TUG, the TUG, the SCT, and JKOM scores. Propensity scores were subsequently used to match participants on a one-to-one basis using the nearest-neighbor method without replacement and a caliper width of 0.2 standard deviations of the logit of the propensity score.

Figure 1 from Kubo et al. (2020)

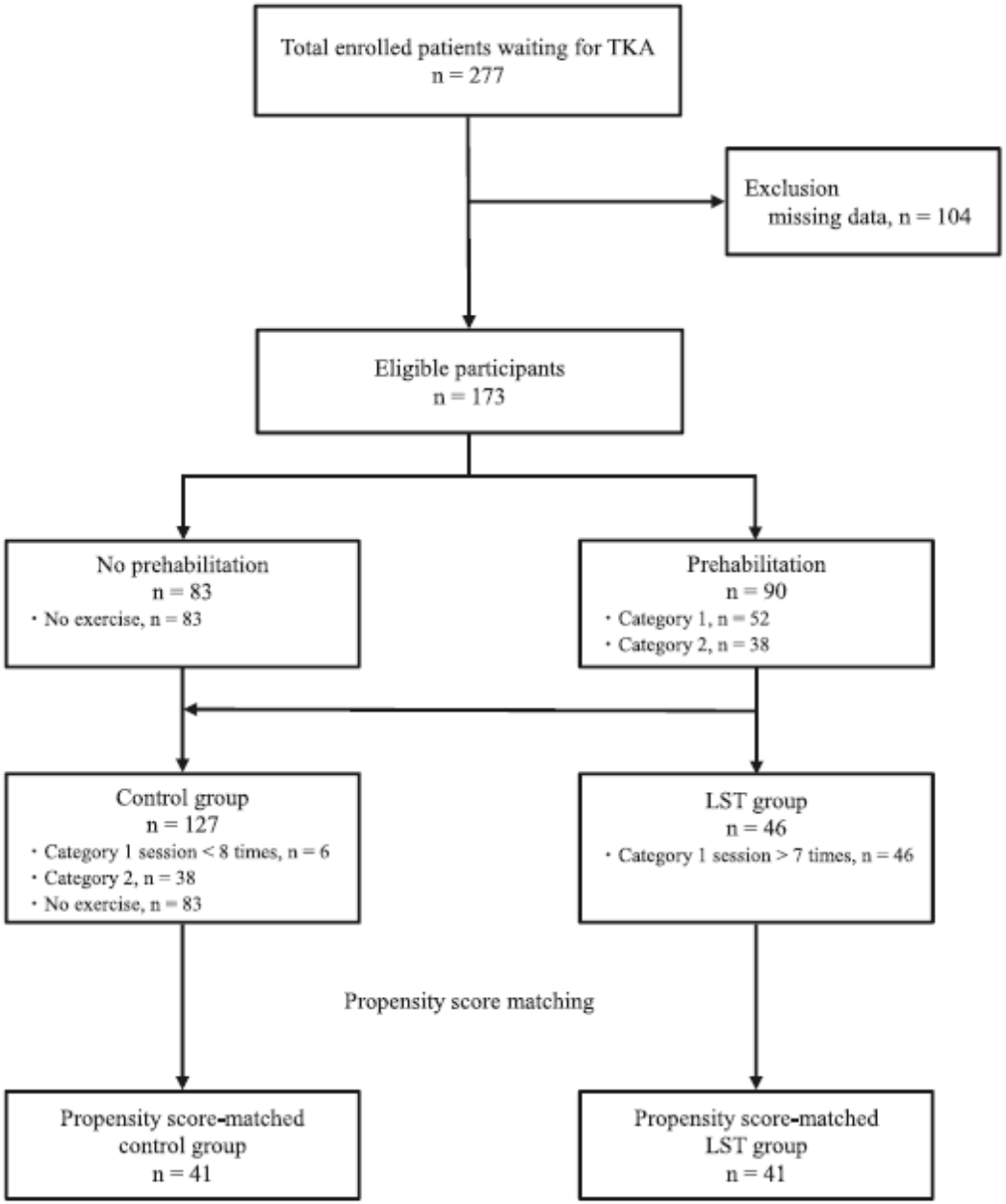


Fig. 1 Study flowchart. The LST group included participants who performed category 1 sessions (LST and aerobic exercise) more than seven

Table 1 from Kubo et al. (2020)

Table 1 Preoperative characteristics of participants and tourniquet time in the control and LST groups

	All participants			Matched participants		
	Control ($n = 127$)	LST ($n = 46$)	SMD	Control ($n = 41$)	LST ($n = 41$)	SMD
Age (years), median (IQR)	74 (68, 79)	71 (66, 75)	0.06	71 (66, 75)	71 (67, 75)	0.00
Male, n (%)	40 (31)	4 (9)	0.59	5 (12)	4 (10)	0.08
BMI (kg/m^2), median (IQR)	26 (23, 28)	25 (23, 28)	0.02	26 (24, 29)	25 (23, 28)	0.00
Current medical history, n (%)						
Heart disease	15 (12)	3 (7)	0.18	2 (5)	3 (7)	0.10
Diabetes	24 (19)	8 (17)	0.04	8 (20)	8 (20)	0.00
Hyperlipidemia	51 (40)	20 (43)	0.07	18 (44)	18 (44)	0.00
Rheumatoid arthritis	1 (1)	2 (4)	0.23	0 (0)	2 (5)	0.32
KL grade 3 of surgical knee, n (%)	14 (12)	5 (11)	0.00	4 (10)	5 (12)	0.08
Contralateral knee, n (%)						
OA and TKA, n (%)	102 (80)	38 (83)	0.06	32 (78)	33 (80)	0.06
Quadriceps strength, median (IQR)	1.3 (1.1, 1.7)	1.4 (1.2, 1.7)	0.05	1.3 (1.1, 1.7)	1.4 (1.2, 1.7)	0.01
T-handle cane usage, n (%)	7 (6)	1 (2)	0.17	0 (0)	1 (2)	0.22
SCT (s), median (IQR)	24 (16, 34)	20 (14, 25)	0.03	21 (17, 31)	22 (16, 27)	0.01
JKOM (points), median (IQR)	37 (25, 49)	32 (25, 45)	0.01	36 (32, 46)	34 (27, 48)	0.00
Tourniquet time (min), median (IQR)	58 (54, 66)	56 (52, 67)	0.01	56, (53, 63)	56 (52, 65)	0.01

Preoperative characteristics and tourniquet time between the groups were compared using standardized mean differences. *Abbreviations*: *LST* low-intensity resistance exercise with slow movement and tonic force generation, *SMD* standardized mean difference, *IQR* interquartile range, *BMI* body mass index, *KL* Kellgren and Lawrence, *OA* osteoarthritis, *TKA* total knee arthroplasty, *SCT* stair climb test, *JKOM* Japanese Knee Osteoarthritis Measure

Study Limitations

There are several limitations that need to be considered. First, the study included a small number of each group participants. Second, this was a single-center retrospective study; accounting for all unmeasured or unknown confounders affecting the outcomes was impossible, even after propensity score matching. Third, some variables remained imbalanced after propensity score matching. However, it is important to note that most imbalanced variables were worse in the LST group than that in the control group, suggesting that preoperative LST program may have improved early QW even in cases with relatively low physical function. Finally, given that QW was assessed only on postoperative day 4, it remains uncertain whether early QW suppression can optimize long-term postoperative recovery. In future, a large-scale multicenter randomized controlled trial with long-term follow up is needed to address these limitations.

Results and Conclusions

Results: Propensity score matching generated 41 matched pairs who had nearly balanced characteristics. The LST group had a significantly lower knee and thigh swelling, QW, and ΔTUG compared to the control group (all, $p < 0.05$). No significant differences in Δknee pain during the QST and TUG were observed between both groups (both, $p > 0.05$).

Abstract Conclusions: The present study demonstrated the beneficial effects of preoperative LST program on knee swelling, thigh swelling, QW, and walking disability immediately after TKA.

Conclusions Section (in the body of the paper)

The present study showed that preoperative LST program exerted beneficial effects on knee and thigh swelling, QW, and walking disability immediately after TKA. Future research addressing the limitations of this study is nonetheless needed to confirm the validity of our findings.

OSIA Discussion

Details to come.

Section 4

Some Extensions to Propensity Matching

Is Regression Adjustment Unnecessary?

- Matching and stratification are old and trusted methods of adjustment for observational studies, but the difficulty of implementing them led earlier practitioners to prefer regression.
- Modern extensions to matching methods let us perform optimal matches, full matches and optimal full matches, and to control imbalance (or at least reduce bias reduction) in ways that have become attainable only in recent years.

Good references include Rosenbaum (2010) and Hansen (2004) for example.

General Approaches to Optimal or Near-Optimal Constrained Matching

1. Calculate propensity scores
2. Establish a **distance matrix**

This is just a table with one row for each treated subject and one column for each potential control.

- The "distances" can be squared differences in propensity scores between the subjects, Mahalanobis distances, or something else.
- To use calipers, we set to ∞ all cells in the table corresponding to a propensity difference which exceeds the caliper.

A Small Distance Matrix

Consider four treated subjects (T1, T2, T3 and T4) and six control subjects (C1, C2, C3, C4, C5 and C6.)

- We have a difference score (perhaps the absolute difference in propensity for treatment) for each comparison. Some of these are infinite.
- We also have each subject categorized as (Y)oung or (O)ld, and we haven't decided yet how important this is for our matching.

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

OK, so Who Gets Matched?

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

- Now, who gets matched?

OK, so Who Gets Matched?

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

- Now, who gets matched?
- Treated subject T1 matches to C1

OK, so Who Gets Matched?

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

- Now, who gets matched?
- Treated subject T1 matches to C1
- T2 matches to C3

OK, so Who Gets Matched?

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

- Now, who gets matched?
- Treated subject T1 matches to C1
- T2 matches to C3
- T3 matches to C4 (or maybe C6 - is age important?)

OK, so Who Gets Matched?

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

- Now, who gets matched?
- Treated subject T1 matches to C1
- T2 matches to C3
- T3 matches to C4 (or maybe C6 - is age important?)
- T4 matches to C6 (or C2, or C4, hmmm....)

Almost Exact Matching

- Suppose a few of the covariates are of enormous importance - want to match exactly on them wherever possible.

We could add a penalty (but perhaps not an infinite penalty) to the distance matrix when the specified covariates fail to match, and that is the main approach that we use.

- Adding 2 to the Mahalanobis distance for mismatches roughly doubles the importance of that covariate as compared to the others, for example.

There's a lot of active work in this area developing various algorithms that permit finer control.

"Fine Balance" in Matching

- Constrain optimal matching that forces a nominal variable to be balanced, without restricting who is matched to whom.

This is especially useful if...

- you have a nominal variable with many levels
- you have a rare binary variable that is difficult to control using a distance
- you are focused on the interaction of several nominal variables

It is also possible to get specific imbalance patterns.

Fine Balance: Initial Distance Matrix

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

Suppose we want to get optimal balance on the propensity score while matching perfectly on the age category (Y/O).

- We have 4 treated subjects (1 young, 3 old)
- We have 6 potential controls (2 young, 4 old)
- So we need to remove 1 young and 1 old in matching

Fine Balance: Augmented Distance Matrix

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24
Extra 1	0	∞	∞	0	∞	∞
Extra 2	∞	0	0	∞	0	0

Add 2 rows to the matrix, then run the match

- *Extra 1* pulls away one young control
- *Extra 2* pulls away one old control

The binary age category will be perfectly balanced across the matched sample, but the partners within each individual pair are not required to be in the same age category.

Fine Balance General Procedure

To get the minimum distance match with fine balance (on a nominal covariate, say GROUP)...

1. Cross tabulate GROUP with treatment indicator
2. Determine # of controls to remove from each category of GROUP to achieve perfect balance
3. Add one row for each control that must be removed, with 0 distance to its own category and infinite distance to all others
4. Find an optimal match for this square matrix
5. Discard extra rows and their matched controls

Section 5

Full Matching

Full Matching in Observational Studies

- In the past, it has been tough to implement full matching in observational studies, even though it is appealing in principle.
- Alignment of comparable treated and control subjects is as good as any alternate method, and potentially much better.
- Hansen (2004) modifies full matching with modifications to minimize variance as well as bias

In this example,

- Optimal full matching removes as much as 99% of the bias along a PS on which treated and control means are separated by 1.1 SD's.
- Reduces to insignificance biases along 27 covariates, while making use of more, not less, of the data than regression based analyses.

Hansen (2004) SAT Coaching Study

- Survey of a random sample of 1995-1996 SAT test takers about their preparation
- 12% of respondents had completed extracurricular test preparation courses
- Matching looked unattractive to the original researchers due to significant reduction in sample size, but they only considered 1:1 matching.
- Do 1:k matching options look better?

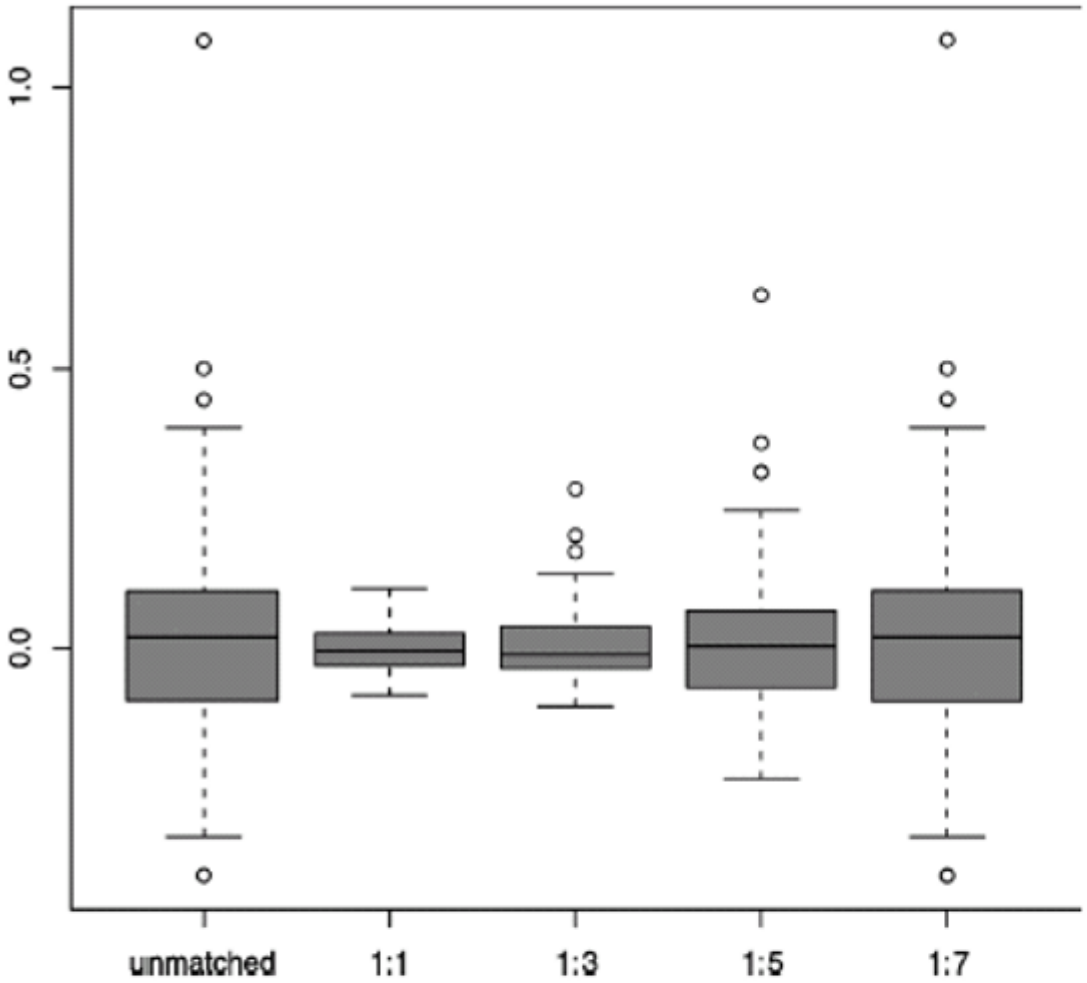


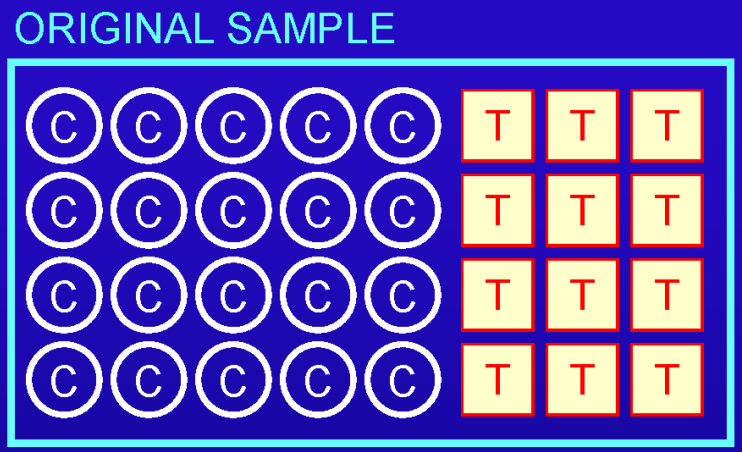
Figure 1. Covariate Imbalances in 1 : k Matching. Each boxplot represents standardized biases in the 99 categories of the 27 categorical covariates along with standardized bias in the propensity score (which in each plot is the uppermost outlier). Strictly speaking, the matching represented at far right is not a 1 : 7 matching but a blend of six 1 : 6 and 494 1 : 7 matched sets.

Covariate Imbalances in 1:k Matching

- In all of these cases, we're using less data
- Still some imbalance

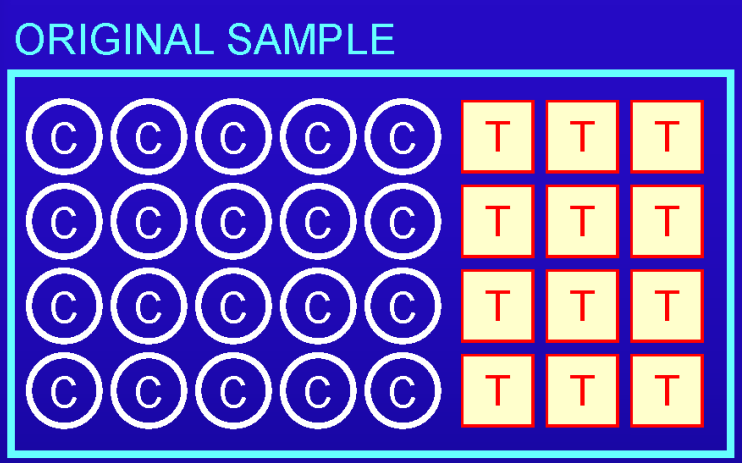
Hansen 2004

Optimal Full Matching



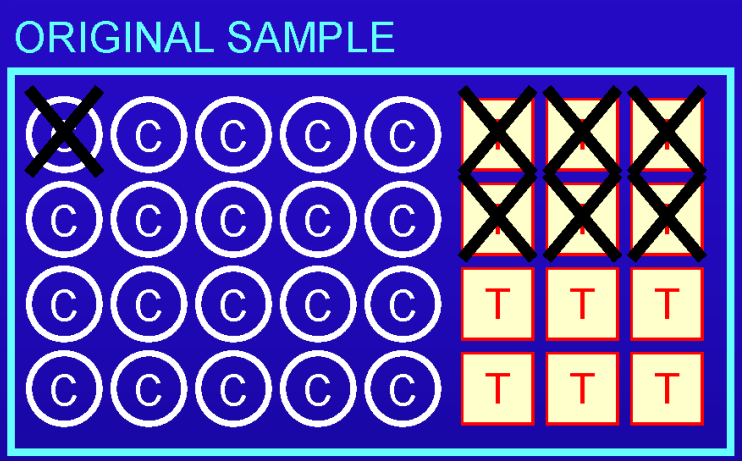
- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).

Optimal Full Matching



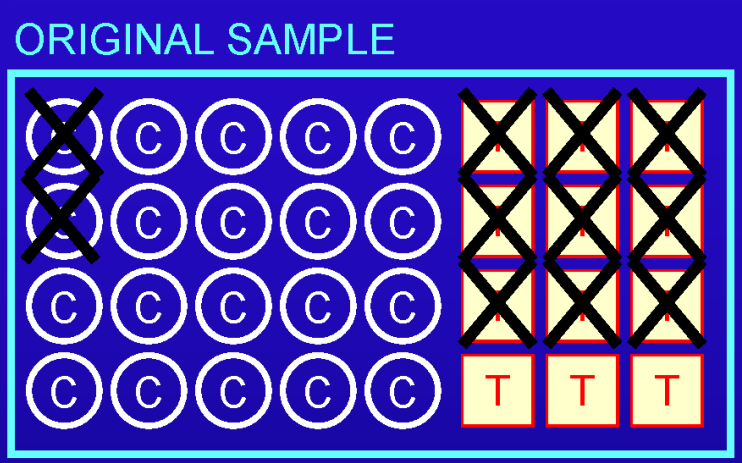
- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

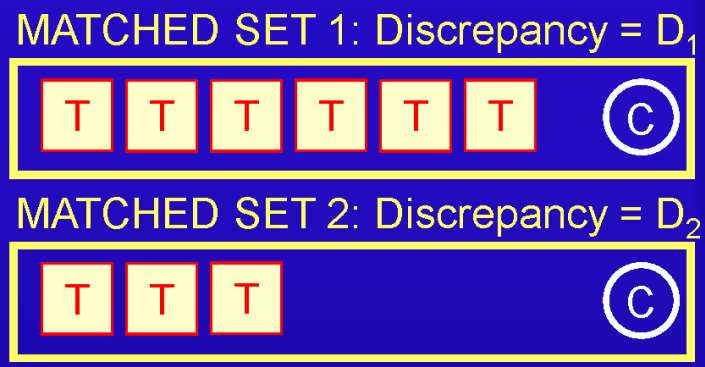
Optimal Full Matching



- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

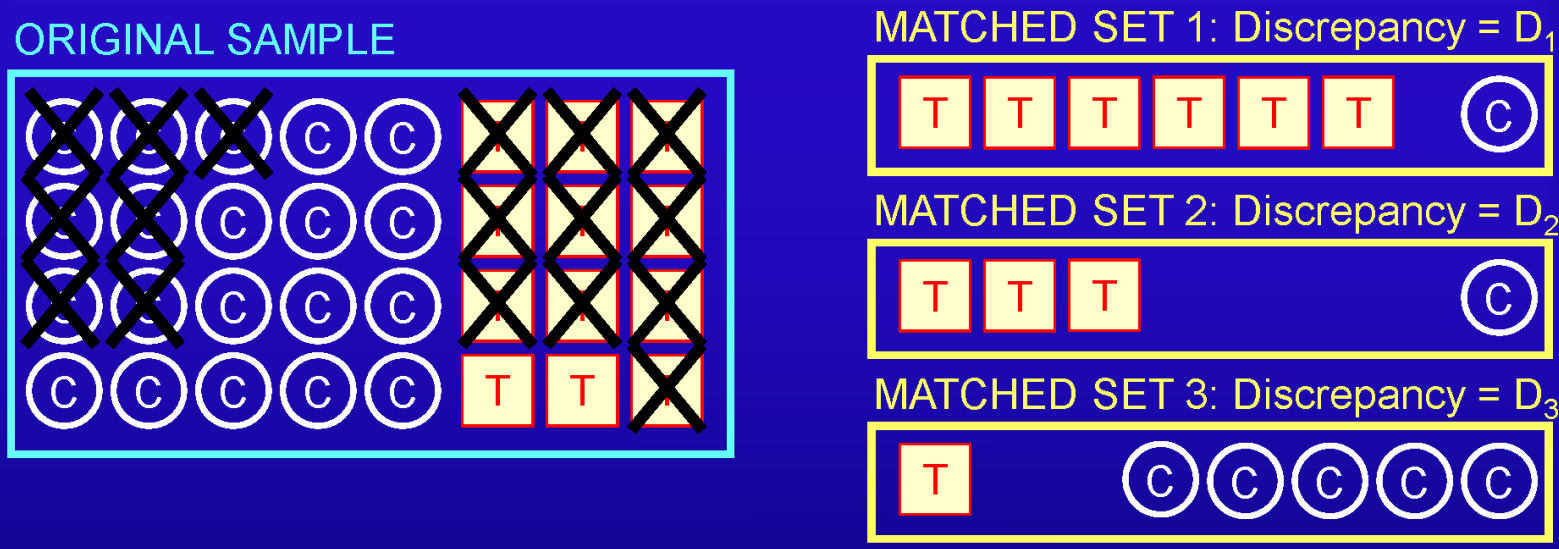
Optimal Full Matching





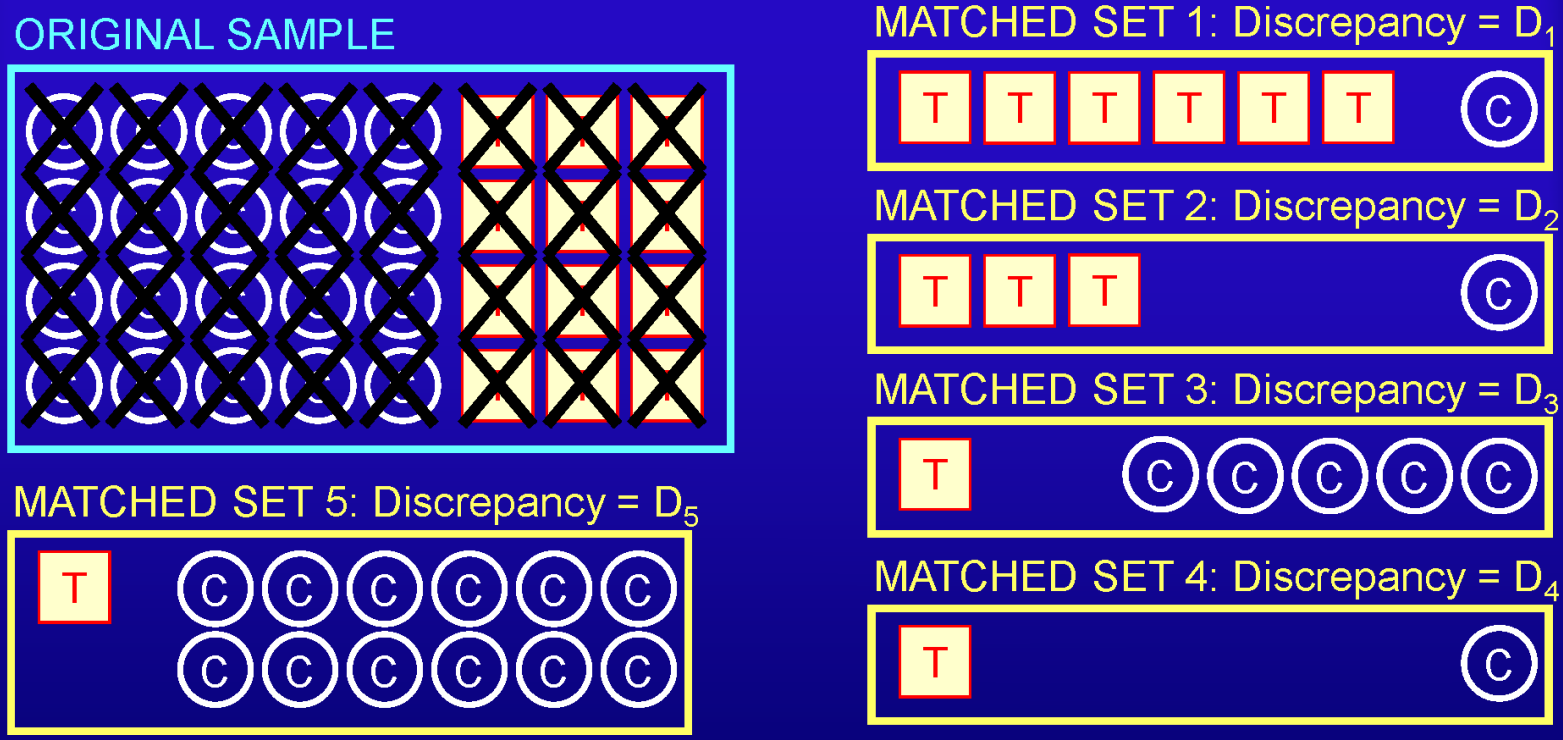
- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

Optimal Full Matching



- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

Optimal Full Matching



- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

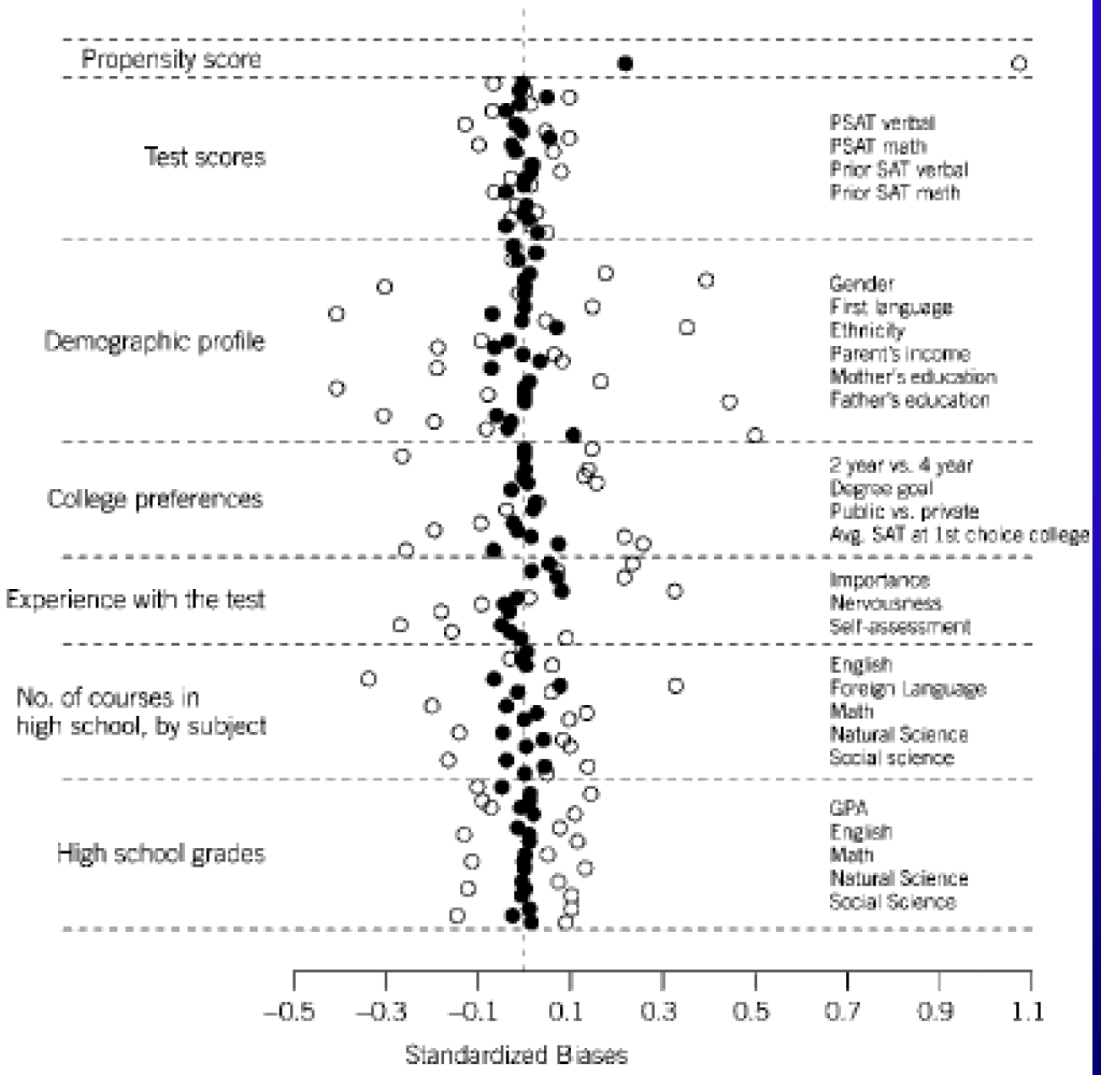


Figure 3. Standardized Biases Without Stratification or Matching, Open Circles, and Under the Optimal [.5, 2] Full Match, Shaded Circles.

Standardized Bias Plot

- Open circles are for standardized biases before matching
- Shaded circles describe results after full match

SAT Coaching Study Results

- Raw differences of treated and control group means were 41 points on Math and 9 on Verbal
- Full matching leads to aggregate contrasts of 26 points on Math and 1 point on the verbal.
  - Standard errors for these estimates are around 5 points.
- Surprised that Verbal effect is so small?
  - Control is not "no prep at all"
  - Estimated effect of treatment on the controls is 3 for Math and -8 on Verbal.
- Method doesn't require homogeneity of coaching effects.
- Whether and to what degree coaching is beneficial appears to vary greatly across students.

Next Class (2024-02-29)

1. Analyses of the SUPPORT / Right Heart Catheterization Study
2. Propensity Scores and Sensitivity Analysis
3. Read Rosenbaum Chapter 6
4. Skim Elbadawi 2021 from our sources

Lab 3 (weighting and double robust analyses) is due at the start of our next class.