500 Class 06 (Recording)

https://thomaselove.github.io/500-2025/

2024-02-20

What's in these Slides?

- Designing Observational Studies (Rubin, 2001)
- Rubin's Rules
  - in the `toy`, `lindner` and `dm2200` examples
- Austin and Mamdani (2006) Example

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

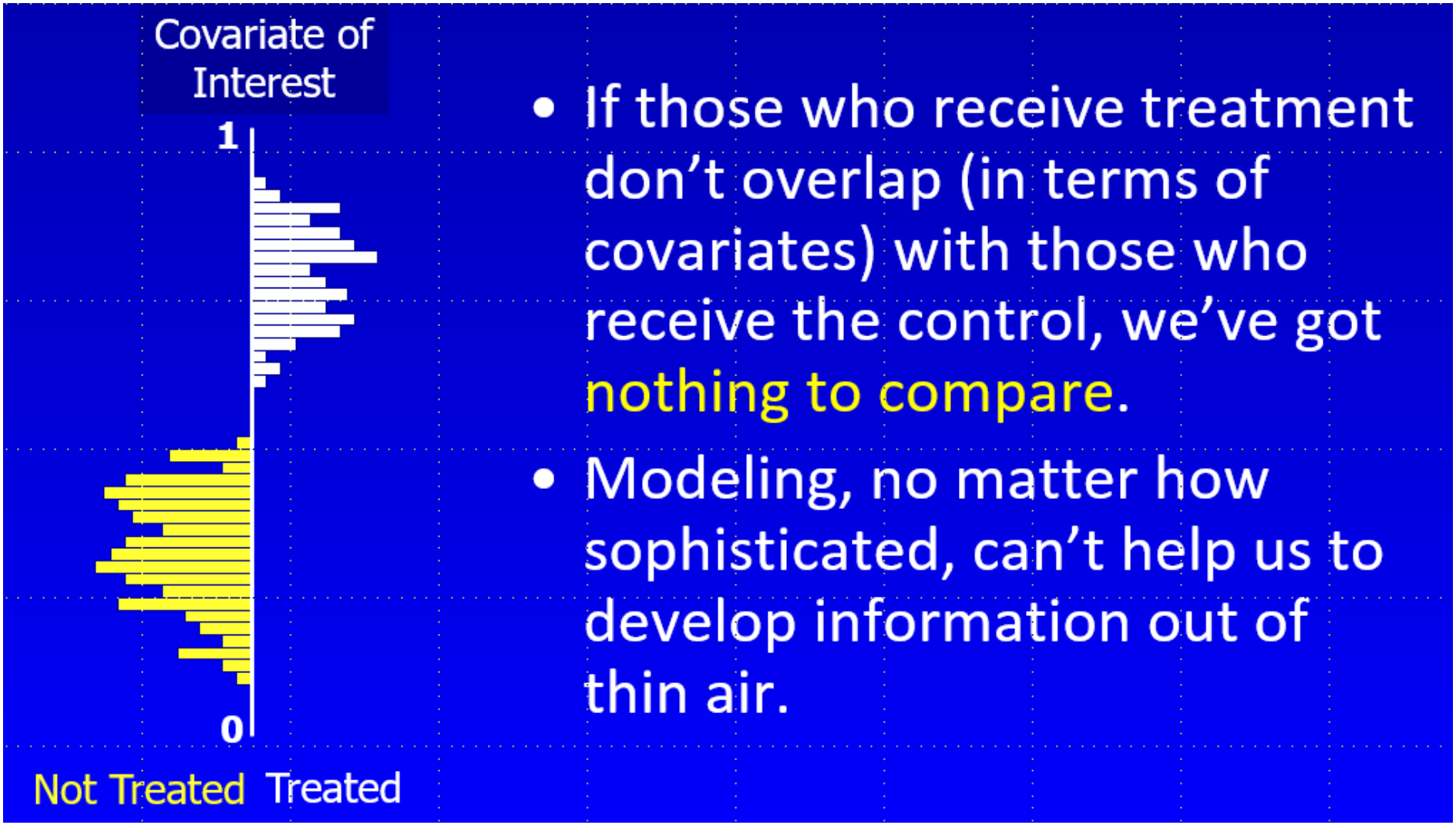
- The Importance of Covariate (PS) Overlap
- How To Check for Overlap Effectively
- Designing Like You're Doing an Experiment
- Using Matching, Subclassification and Weighting
- Propensity Scores are "Fair Game" - No Outcomes!

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.

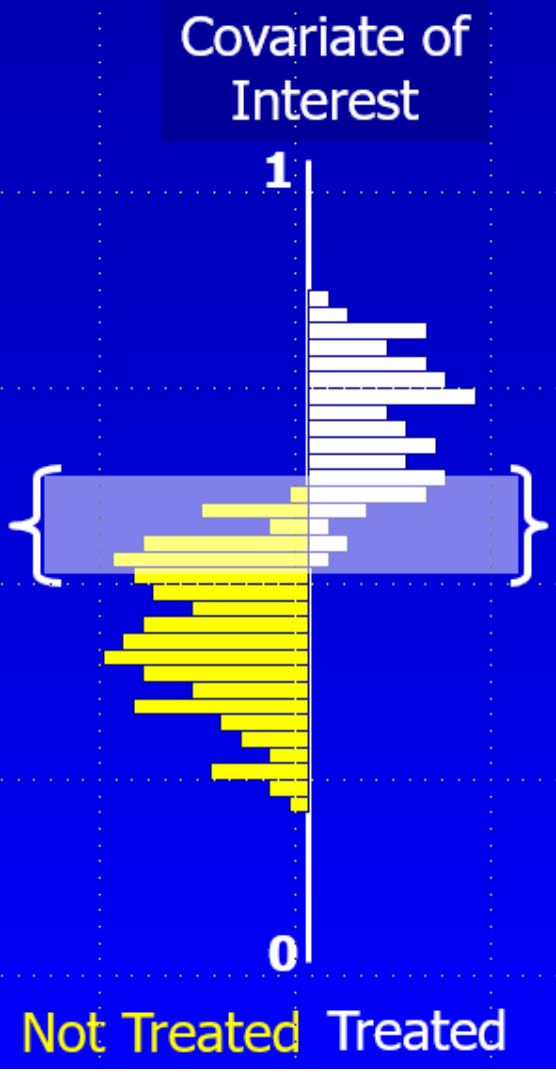
Goal: Use propensity scores to assemble treatment groups that have comparable distributions on all measured covariates.

Issue 1: Overlap

How much overlap in the covariates do we want?



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.

Rubin's Rules in the `toy` example

Section 7. Rubin's Rules to Check Overlap **Before** Propensity Adjustment

1. Difference in mean (linps) = 85.9% of a pooled standard deviation
2. Variance of linps ratio = 0.63
3. Dot chart of the residual variance ratios in section 7.3.1.

Rubin's Rules: `toy` example

Section 17.3. Quality of Balance by Rubin's Rules after adjustments...

Approach	Rule 1	Rule 2	Rule 3
"Goal"	0 to 50	0.5 to 2	0.5 to 2
Unadjusted	85.9	0.63	0.72 to 1.76
1:1 Match	25.3	1.24	0.8 to 1.3
ATT Weighting	-9.1	0.79	Not evaluated

Clearly, the matching and propensity weighting show improvement over the initial (no adjustments) results, although neither is completely satisfactory in terms of all covariates. In practice, I would be comfortable with either a 1:1 match or a weighting approach.

Rubin's Rules in the `lindner` example

Section 8. Rubin's Rules to Check Overlap **Before** Propensity Adjustment

1. Difference in mean (linps) = 61.9% of a pooled standard deviation
2. Variance of linps ratio = 1.67
3. rule 3 not evaluated in this example

Rubin's Rules: `lindner` example

Quality of Balance by Rubin's Rules after adjustments...

Approach	Rule 1	Rule 2
"Goal"	0 to 50	0.5 to 2
Unadjusted	61.9	1.67
1:1 Match w/o repl.	65.5	1.78
1:1 Match w/repl.	0.9	1.05
ATT Weighting	4.8	0.90

Rubin's Rules: dm2200 example

Quality of Balance by Rubin's Rules after adjustments...

Approach	Rule 1	Rule 2
"Goal"	0 to 50	0.5 to 2
Unmatched	206.5	0.75
1:1 greedy match	24.6	1.86
1:2 greedy match	50.0	2.44
1:3 with replacement	3.1	1.12
Caliper (0.2) 1:1 match	0.1	1.02
1:1 optimal match	24.6	1.86

Section 2

Austin and Mamdani (2006)

A case study

STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:2084–2106
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2328

A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use

Peter C. Austin[1,2,3,*,†] and Muhammad M. Mamdani[1,3,4]

[1]*Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*
[2]*Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada*
[3]*Department of Health Policy, Management and Evaluation, University of Toronto, Canada*
[4]*Faculty of Pharmacy, University of Toronto, Canada*

Summary (excerpt)

There is an increasing interest in the use of propensity score methods to estimate causal effects in observational studies. However, recent systematic reviews have demonstrated that ***propensity score methods are inconsistently used and frequently poorly applied*** *in the medical literature.*

In this study, we compared the following propensity score methods for estimating the reduction in all-cause mortality due to statin therapy for patients hospitalized with acute myocardial infarction: propensity-score matching, stratification using the propensity score, covariate adjustment using the propensity score, and weighting using the propensity score.

Introduction (1/2)

There is an increasing interest in using observational data to assess the impact of medical treatment or therapy on health outcomes.

There is a growing interest in the use of propensity score-based methods for estimating treatment effects in observational studies.

Three propensity score-based methods are commonly used in the medical literature: matching, covariate adjustment, and stratification or subclassification.... There is no consensus in the clinical literature as to which method is preferable. Furthermore, there is only a limited awareness of the relative strengths and limitations of each propensity score method.

Introduction (2/2)

The purpose of the current study was two-fold. First, to compare estimates of treatment effectiveness obtained using different propensity score methods. In doing so, the relative merits and limitations of each method will be highlighted. Second, to carry out a detailed propensity score analysis, which will serve as a model for clinical investigators who wish to implement propensity score methods.

As a test case, we examine the effect of statin lipid-lowering therapy on reducing all-cause mortality for patients discharged alive from hospital with a diagnosis of acute myocardial infarction (AMI).

Data Sources

Detailed clinical data were collected on a sample of 11 524 patients discharged from Ontario hospitals between April 1, 1999 and March 31, 2001 by retrospective chart review.

Data on patient history, cardiac risk factors, comorbid conditions and vascular history, vital signs, and laboratory tests were collected for this sample. Furthermore, data on medications prescribed at discharge were available for each patient. Linking patients to the registered persons database (RPDB) using encrypted health card numbers allowed us to determine each patient's vital status. We allowed each patient to have 3 years of follow-up post-discharge. Patients who were missing data on important vital signs at admission or laboratory values were excluded from all subsequent analyses.

Assessing Balance

Differences in measured characteristics between treated and untreated patients were assessed using two methods.

First, the statistical significance of the difference in either the proportion of patients having a dichotomous risk factor or the difference in the mean of a continuous covariate between treated and untreated patients was tested. A chi-square test was used for dichotomous variables and a t-test was used for continuous variables.

Second, the standardized difference was computed for each covariate. It has been suggested that a standardized difference of greater than 10 per cent represents meaningful imbalance in a given covariate between treatment groups.

Potential Confounders

Clinical data contained the following variables that were potential confounders of the treatment effect: demographic characteristics: age, gender; presenting signs and symptoms: shock and acute congestive heart failure (acute CHF)=pulmonary oedema; classical cardiac risk factors: diabetes, CVA=TIA (history of cerebrovascular accident or transient ischaemic attack), history of hyperlipidaemia, hypertension, family history of heart disease, and smoking history; comorbid conditions: angina, cancer, congestive heart failure (CHF)=pulmonary oedema, and renal disease; vital signs on admission: systolic and diastolic blood pressure, heart rate, and respiratory rate; laboratory test results (haematology): haemoglobin (Hgb), white blood count (WBC); and laboratory test results (chemistry): sodium, potassium, glucose, and creatinine.

Table I. Comparisons of statin users and non-users.

Characteristic	Statin non-users $N = 6055$	Statin users $N = 3049$	P-Value
Demographic characteristics			
Age	68.11 ± 13.85	63.36 ± 12.39	<0.001
Female	2241 (37.0%)	887 (29.1%)	<0.001
Presenting characteristics			
Shock	46 (0.8%)	12 (0.4%)	0.038
Acute CHF/pulmonary oedema	316 (5.2%)	122 (4.0%)	0.010
AMI risk factors			
Family history of CAD	1762 (29.1%)	1177 (38.6%)	<0.001
Diabetes	1561 (25.8%)	774 (25.4%)	0.684
CVA/TIA	610 (10.1%)	237 (7.8%)	<0.001
Hyperlipidaemia	1138 (18.8%)	1761 (57.8%)	<0.001
High BP	2681 (44.3%)	1453 (47.7%)	0.002
Current smoker	2004 (33.1%)	1070 (35.1%)	0.057
Comorbidities			
Angina	1869 (30.9%)	1086 (35.6%)	<0.001
Cancer	191 (3.2%)	73 (2.4%)	0.041
CHF	275 (4.5%)	91 (3.0%)	<0.001
Renal disease	34 (0.6%)	13 (0.4%)	0.396
Vital signs on admission			
Systolic BP	148.69 ± 31.57	149.35 ± 30.07	0.338
Diastolic BP	83.62 ± 18.62	84.49 ± 18.00	0.033
Heart rate	84.60 ± 24.31	81.71 ± 22.96	<0.001
Respiratory rate	21.20 ± 5.74	20.30 ± 4.78	<0.001
Laboratory values			
White blood count	10.34 ± 4.87	10.03 ± 4.42	0.003
Haemoglobin	137.54 ± 19.35	140.64 ± 16.92	<0.001
Sodium	138.92 ± 3.92	139.22 ± 3.29	<0.001
Glucose	9.43 ± 5.11	9.24 ± 5.30	0.092
Potassium	4.10 ± 0.57	4.07 ± 0.51	0.006
Creatinine	105.71 ± 65.45	99.89 ± 50.03	<0.001

Note: Continuous variables are reported as mean±standard deviation, while dichotomous variables are reported as number with condition (per cent).

Propensity Score Development (1/2)

We developed a propensity score model to predict the probability that the patient would be given a prescription for a statin at hospital discharge.

The propensity score model was developed to balance the distribution of the possible confounders between treated and untreated patients within each quintile of the estimated propensity score.

The iterative algorithm for developing the propensity score allowed for the inclusion of main effects, interaction terms, in addition to quadratic and cubic terms for continuous variables.

Propensity Score Development (2/2)

Following the derivation of the propensity score model, untreated patients who had a lower estimated propensity score than any treated patient were excluded from subsequent analyses.

The coefficients of the propensity model were then re-estimated on this reduced data set.

This decision was made because of the belief that the excluded patients were different from all treated patients, and may thus have not been candidates for statin therapy.

Cohort Characteristics

Detailed clinical data were available on 11 524 patients admitted with a diagnosis of AMI. Encrypted health card numbers were missing for four records, which were excluded since linkage to the RPDB was not possible without a valid health card number. A further 1137 patients died during the index hospitalization, resulting in a cohort of 10 383 patients discharged alive from an index hospitalization for AMI. Of these, a further 1279 patients were removed from the analyses due to missing data on important confounding factors (vital signs on admission and laboratory test results). This resulted in a final sample size of 9104 AMI survivors who were used for the development of the propensity score model.

Overall, 3049 (33.5%) patients received a statin prescription at discharge. The 3-year mortality rate was 14.2% and 25.3% for those who did and did not receive a prescription for a statin at discharge, respectively.

The propensity score model

The final propensity score model included 256 variables: 27 main effects and 229 two-way interactions.

Following the derivation of the propensity score model, 95 untreated patients had a lower estimated propensity score than any treated patient, and were excluded from all subsequent analyses. Thus, 9009 patients were used in all subsequent analyses.

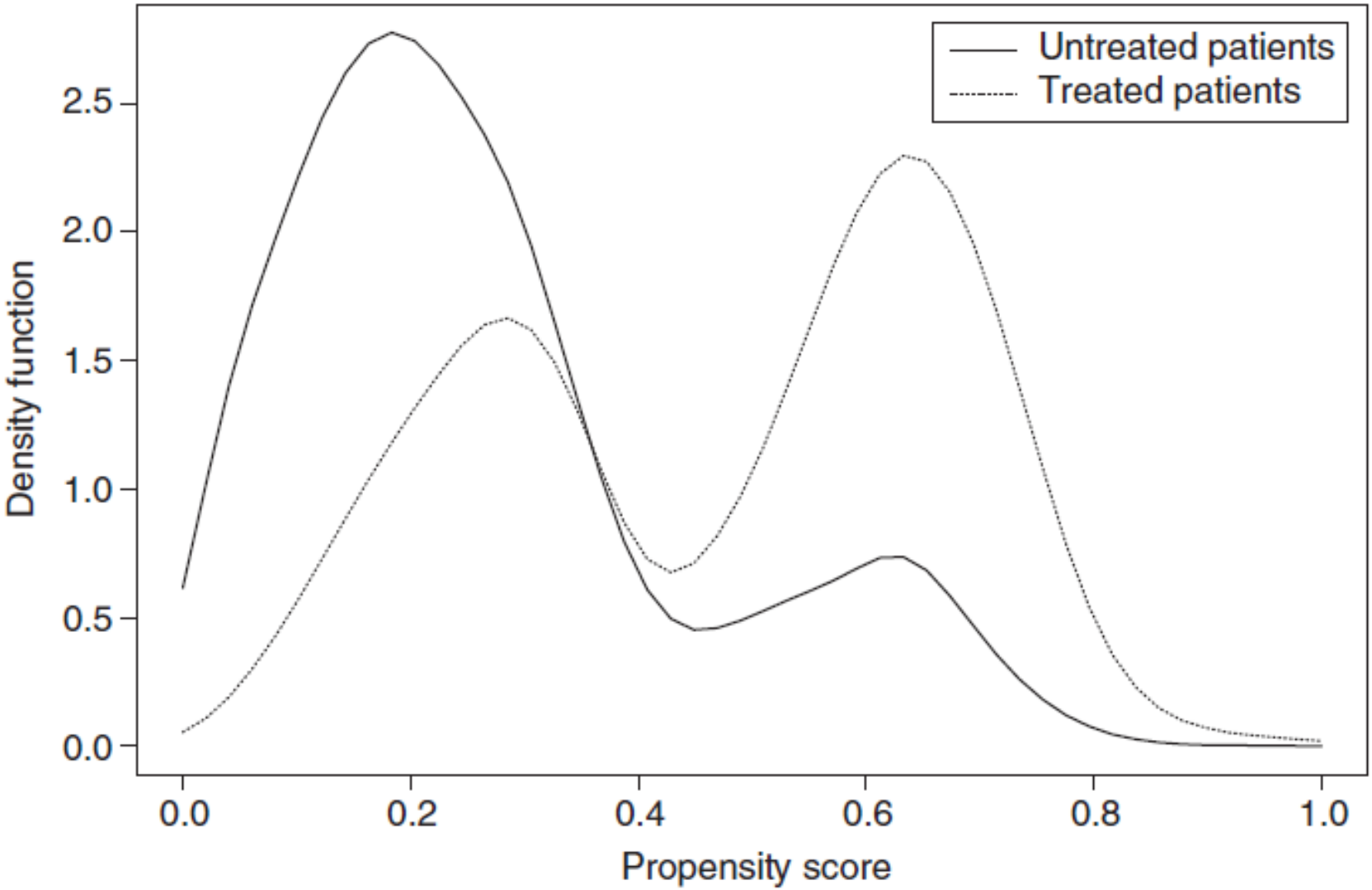


Figure 1. Distribution of propensity score in treated/untreated patients.

Using the Propensity Score

Once the propensity score model had been developed, we sought to estimate the reduction in all-cause mortality attributable to the post-discharge use of statins. This was done in four different fashions:

- stratifying on the quintiles of the estimated propensity score,
- matching treated and untreated patients using the estimated propensity score,
- covariate adjustment using the estimated propensity score, and
- weighting using the propensity score.

Table II. Standardized differences between treated and untreated patients.

Variable	Unmatched sample	Matched sample	Stratified analysis				
			Q1	Q2	Q3	Q4	Q5
Diabetes	−0.9	−0.2	9.2	−8.1	0.4	8.9	4.4
Renal disease	−1.9	0.0	−2.4	1.9	−3.9	2.9	2.0
Systolic BP	2.1	−0.4	11.9	−4.7	4.6	−5.0	−0.2
Glucose	−3.7	1.5	11.1	−11.8	0.9	2.5	3.5
Smoker	4.2	0.6	−5.4	0.3	4.5	3.2	−9.6
Cancer	−4.6	0.0	−2.1	6.0	0.0	0.3	−4.1
Diastolic BP	4.8	−0.2	5.6	−2.7	6.0	−6.3	−1.6
Acute shock	−4.8	0.0	−10.5	−0.1	−3.5	2.8	8.1
Acute CHF/pulmonary oedema	−5.8	−0.5	−7.8	−5.8	5.0	−4.0	8.9
Potassium	−6.2	−0.6	−10.8	−0.8	−5.8	2.7	−0.9
WBC	−6.7	0.5	−8.4	−2.4	3.9	−3.4	5.6
High BP	6.8	−2.6	1.2	−6.0	2.5	4.8	7.1
CVA/TIA	−8.1	−0.3	−1.5	−7.1	−5.0	1.1	6.7
Sodium	8.1	−0.8	−5.0	11.6	0.8	3.6	−6.6
Congestive heart failure	−8.2	−1.0	−8.5	−3.7	−2.5	−4.8	3.7
Creatinine	−10.0	1.5	−10.3	−2.6	−3.1	2.2	8.1
Angina	10.1	2.1	12.6	−2.2	−3.2	−1.2	7.2
Heart rate	−12.2	−0.2	2.1	−11.4	−0.2	14.2	−1.1
Female	−16.9	2.5	−1.8	−1.3	−6.3	7.7	−5.8
Haemoglobin	17.1	−3.6	4.6	6.3	8.1	−8.0	−3.4
Respiratory rate	−17.2	−1.3	−10.0	−0.4	−3.1	4.1	0.5
Family history	20.2	−1.3	5.8	4.7	0.5	−2.4	−0.7
Age	−36.1	−0.2	−19.7	1.7	−6.0	5.4	1.8
Hyperlipidaemia	87.5	−0.3	−11.6	−5.3	−1.3	30.0	−4.8

Note: Each cell is the per cent standardized difference between treated and untreated patients.

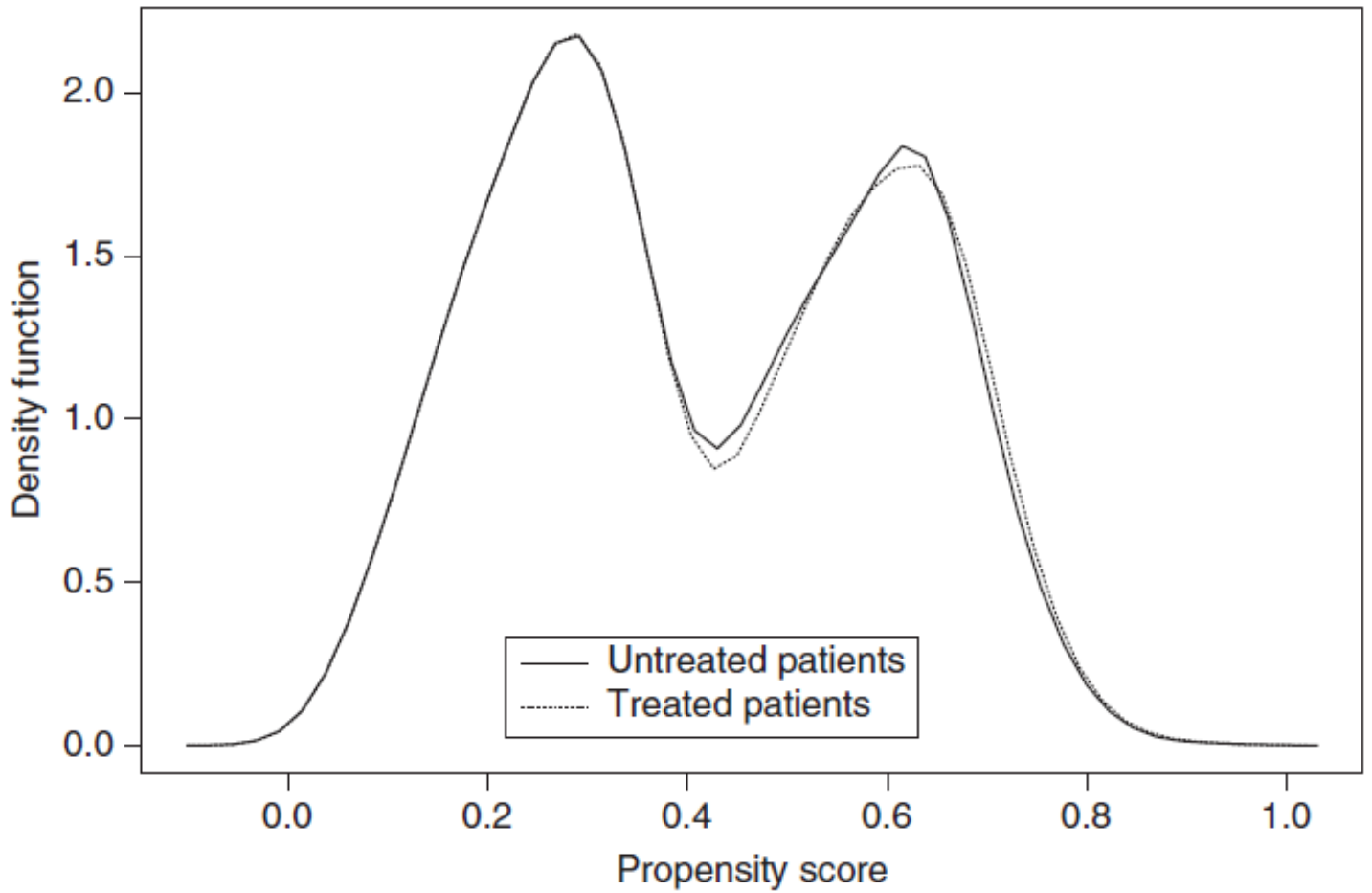


Figure 4. Distribution of propensity score in treated/untreated patients (matched analysis).

Table IV. Comparison of treated and untreated patients in matched sample.

Characteristic	Statin: No $N = 2348$	Statin: Yes $N = 2348$	P-Value
Demographic characteristics			
Age	63.33±12.55	63.30±12.53	0.951
Female	699 (29.8%)	726 (30.9%)	0.392
Presenting characteristics			
Shock	8 (0.3%)	8 (0.3%)	1.000
Acute CHF/pulmonary oedema	85 (3.6%)	83 (3.5%)	0.873
AMI risk factors			
Family history of CAD	919 (39.1%)	904 (38.5%)	0.642
Diabetes	579 (24.7%)	577 (24.6%)	0.944
CVA/TIA	158 (6.7%)	156 (6.6%)	0.906
Hyperlipidaemia	1071 (45.6%)	1068 (45.5%)	0.778
High BP	1107 (47.1%)	1077 (45.9%)	0.364
Current smoker	879 (37.4%)	886 (37.7%)	0.831
Comorbidities			
Angina	735 (31.3%)	758 (32.3%)	0.457
Cancer	58 (2.5%)	58 (2.5%)	1.000
CHF	68 (2.9%)	64 (2.7%)	0.728
Renal disease	7 (0.3%)	7 (0.3%)	1.000
Vital signs on admission			
Systolic BP	149.97±30.07	149.86±29.98	0.903
Diastolic BP	84.84±17.86	84.80±18.10	0.941
Heart rate	81.05±22.60	81.00±22.34	0.943
Respiratory rate	20.26±4.60	20.20±4.52	0.649
Laboratory values			
White blood count	10.12±3.66	10.14±3.94	0.870
Haemoglobin	141.70±17.54	141.08±16.92	0.203
Sodium	139.21±3.82	139.18±3.28	0.769
Glucose	9.18±4.98	9.26±5.43	0.608
Potassium	4.07±0.57	4.07±0.52	0.847
Creatinine	98.05±41.47	98.69±44.54	0.613

Note: Continuous variables are reported as mean±standard deviation, while dichotomous variables are reported as number with condition (per cent).

Table V. Treatment effect for Statin at discharge—adjusted estimates.

Method	Odds ratio	95 per cent confidence interval	*P*-Value
Crude	0.49	(0.44, 0.55)	<0.0001
Propensity score methods			
Stratifying on PS quintiles	0.77	(0.66, 0.89)	0.0003
Stratifying on PS quintiles—within stratum regression adjustment	0.76	(0.65, 0.89)	0.0007
Covariate adjustment—linear term for PS	0.84	(0.74, 0.96)	0.0099
Covariate adjustment—quadratic terms for PS	0.81	(0.70, 0.93)	0.0033
Covariate adjustment—cubic spline for PS	0.81	(0.70, 0.93)	0.0028
Covariate adjustment—linear term for PS with adjustment for additional confounders	0.80	(0.69, 0.93)	0.0038
PS Matching	0.85	(0.72, 0.99)	0.0372
PS weighting—simple model	0.77	(0.64, 0.92)	0.0041
PS weighting—complex model	0.76	(0.63, 0.90)	0.0022
Direct regression adjustment			
Regression adjustment—simple predictive model	0.75	(0.65, 0.85)	<0.0001
Regression adjustment—backwards selection method	0.73	(0.64, 0.84)	<0.0001
Regression adjustment—complex predictive model	0.78	(0.67, 0.91)	0.0014

Conclusions

In conclusion, we compared the estimated reduction in mortality due to receiving a statin prescription at hospital discharge for an AMI using different propensity score methods on a single data set. We demonstrated the breadth of propensity score methods and that these methods allow the estimation of both adjusted as well as absolute and relative treatment effects.

Earlier research has demonstrated that propensity score methods are often poorly implemented in the medical literature. Our intention, through carrying out an extended case study, was to demonstrate the breadth of the propensity score methods and how they can be more fully employed in clinical research.

Zoom Meeting Thursday 2025-02-20 at 10 AM

- How did Lab 2 go?
- OSIA Selections
- Kubo et al. (2020) as an example for OSIA
- Some Extensions to Propensity Matching