500 Class 03

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

2024-02-01

Today's Agenda

- Estimating the Propensity Score (Building the Propensity Model)
- A schematic for propensity matching in a "simple" study
  - Distinguishing ATT and ATE estimates
- Mechanics of Propensity Matching
  - Gum et al. Aspirin use and mortality example
  - Standardized Differences and the Love Plot
  - Incomplete vs. Inexact Matching
- Schematics for other Propensity Methods in "simple" studies
  - Direct (regression) adjustment for the Propensity Score
  - Subclassification / stratification on the Propensity Score
  - Weighting on the Propensity Score
- The SUPPORT / Right Heart Catheterization Study
- Lab 1 (How did it go?)
- A little bit of OSIA and Proposal Advice

Today: No R code. **Class 04**: Nothing but R code

Section 1

Building the Propensity Model

The Propensity Score

$$PS = Pr(\text{received exposure}|\mathbf{X})$$

The propensity score is...

- the conditional probability of receiving the exposure given a particular set of covariates
- a way of projecting meaningful covariate information for a given subject into a single composite summary score in (0, 1)
- a tool that lets us account for *overt* selection bias (things contained in **X**) but not (directly) for the potential biasing effects of omitted/hidden covariates
- often, but not inevitably, fit with a "kitchen sink" logistic regression[1]

$$ln(\frac{PS}{1-PS}) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

[1] See McCaffrey et al 2004 on boosting, and see Brookhart 2006 on variable selection.

What To Include in the Propensity Score Model

- **All** covariates that subject matter experts (and subjects) judge to be important when selecting treatments.
- **All** covariates that relate to treatment and outcome, certainly including any covariate that improves the prediction of treatment group.
- Sop up as much "signal" as possible.

Propensity Score Models: What to Worry About...

1. Do you have a reasonable sample size to build a logistic regression model, e.g., at least 96 subjects + some function of the number of candidate predictors[2]?
2. Is your logistic regression model parsimonious?
3. Are your predictors correlated with one another?
4. Are your predictors statistically significant?
5. Have you performed appropriate diagnostic checks?
6. Have you done bootstrap analyses to assess shrinkage?
7. Have you used cross-validation to aid in model selection?
8. Have you done external validation of your model on new data?
9. Does an ROC-curve analysis suggest your model does well in terms of rank-order discrimination?
10. Have you determined that your model's predictions are well-calibrated?

[2] see Frank Harrell BBR Course Notes

What to **Actually** Worry About

None of those things.

Instead, we simply ensure that the fitted propensity scores (when used in matching, weighting, etc.) adequately balance the distribution of covariates across the exposure groups.

Again, we want to wind up with a **fair basis for comparison** between exposed and control subjects.

Which group is the "exposed" group?

The definition of which group is the "exposed" group (so that it has generally higher propensity scores) and which is the "control" group matters analytically, although it's essentially an arbitrary selection.

- You will make your life easier for our purposes in developing the project by making your "exposed" group the smaller of the two groups in terms of sample size, if possible.

What about Propensity Model Diagnostics?

Rubin (2004) describes "confusion between two kinds of statistical diagnostics"

1. Diagnostics for the successful prediction of probabilities and parameter estimates underlying those probabilities.
2. Diagnostics for the successful design of observational studies based on estimated propensity scores.

Basically, the set of tasks in 1 are irrelevant to 2.

Should we be checking propensity model goodness of fit?

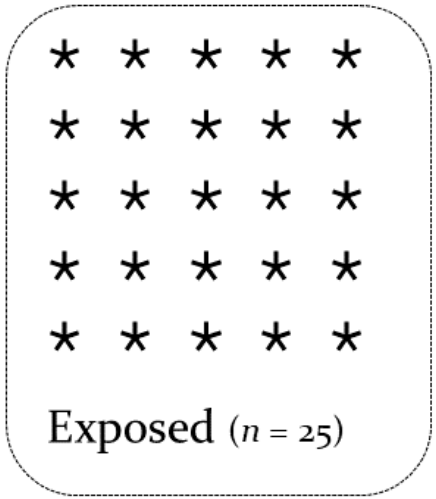
Weitzen et al. (2004): Are tests used to evaluate logistic model fit and discrimination helpful in detecting the omission of an important confounder?

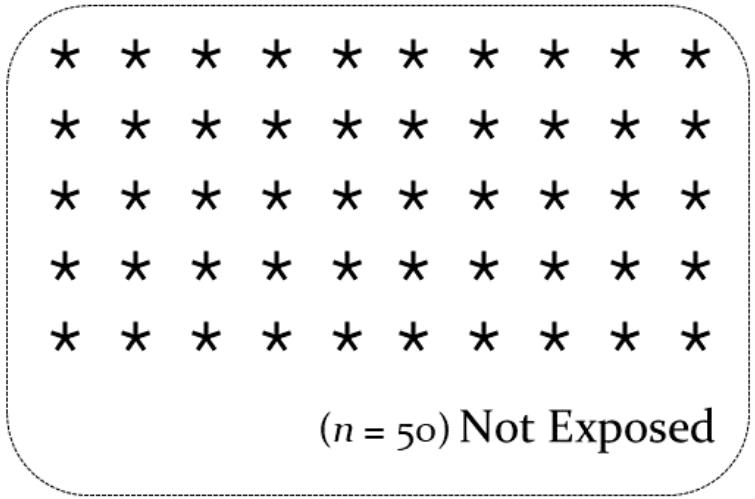
- Simulated data including an important binary confounder, and they compared inclusion to exclusion
- Hosmer-Lemeshow GOF test and C statistic were of no value in detecting residual confounding in treatment effect estimates

Section 2

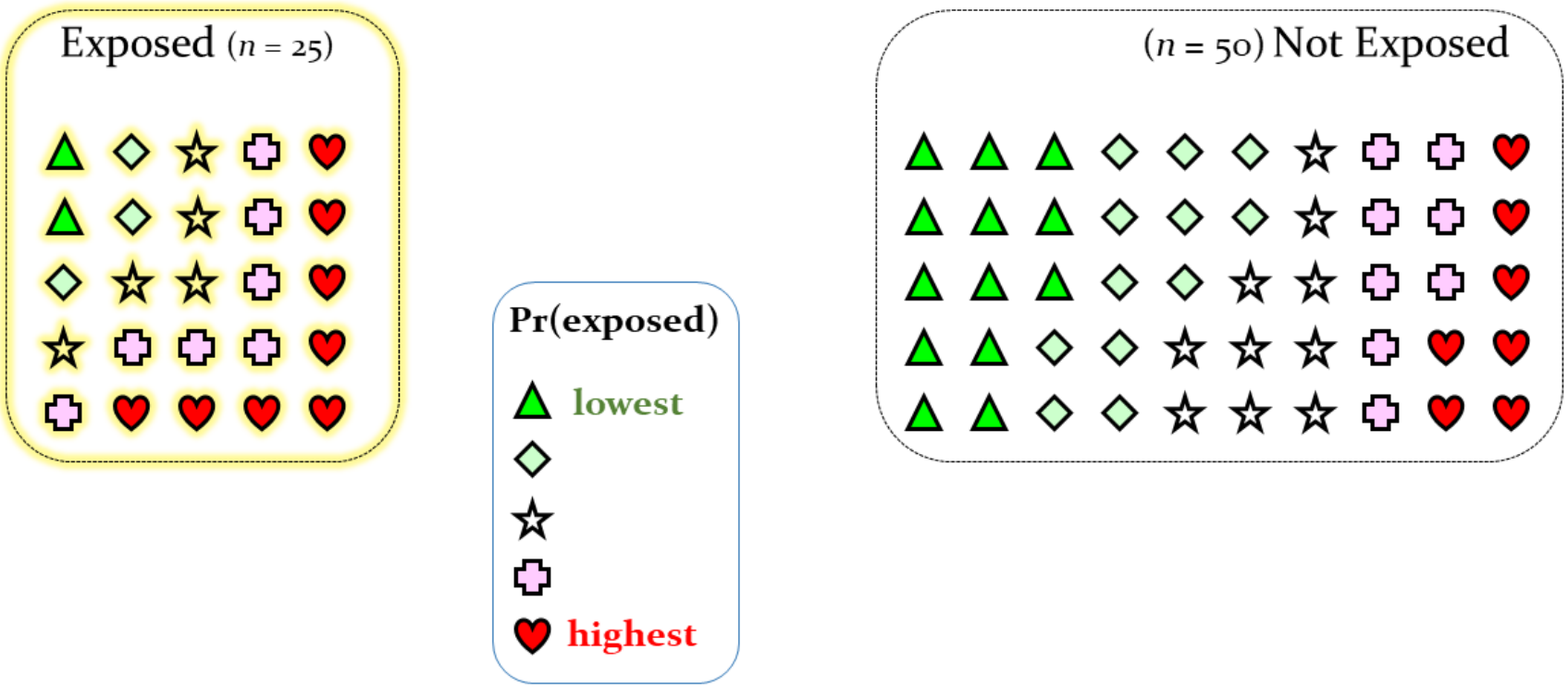
Propensity Matching in A Simple Observational Study

Simple Observational Study

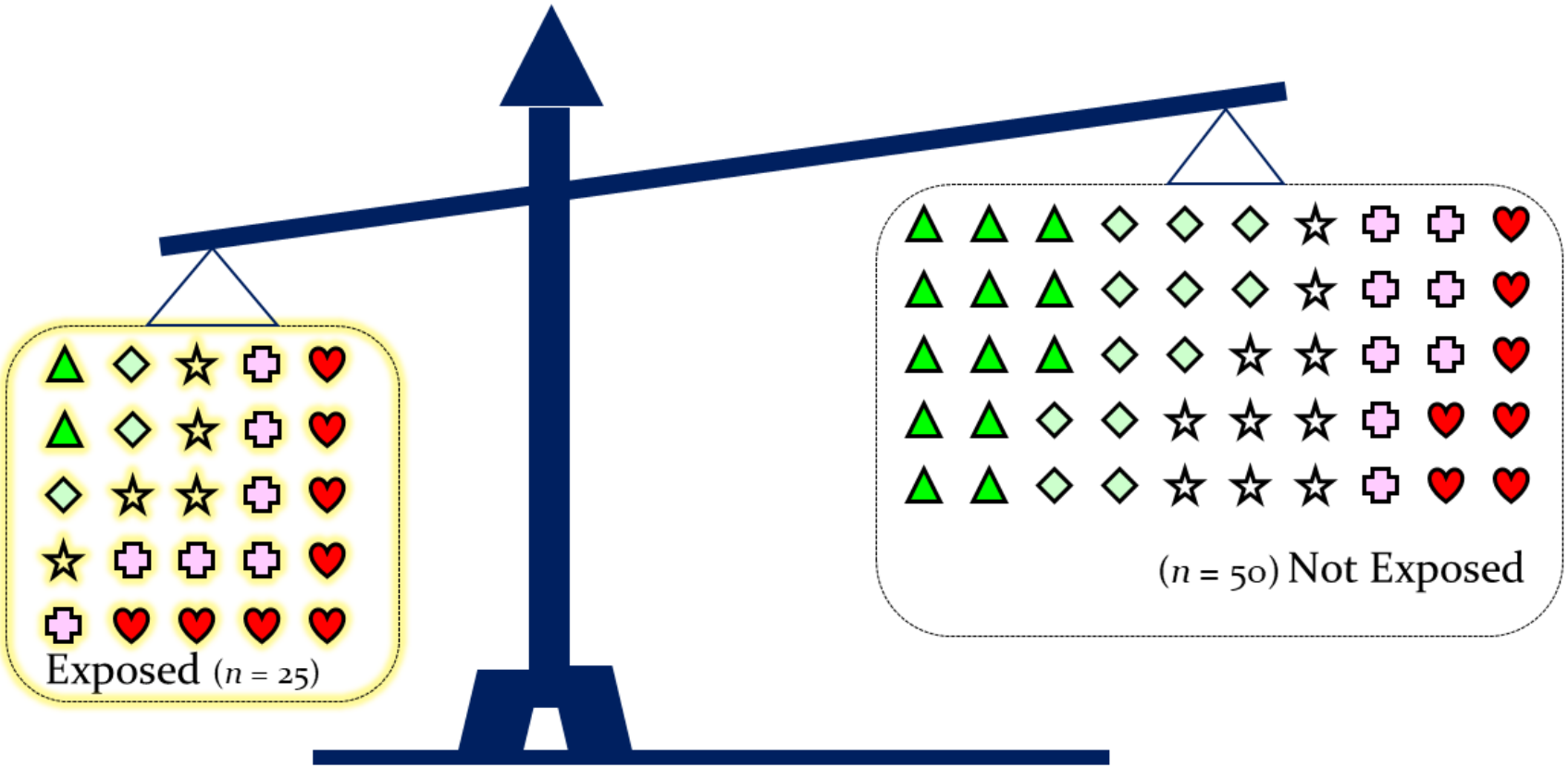




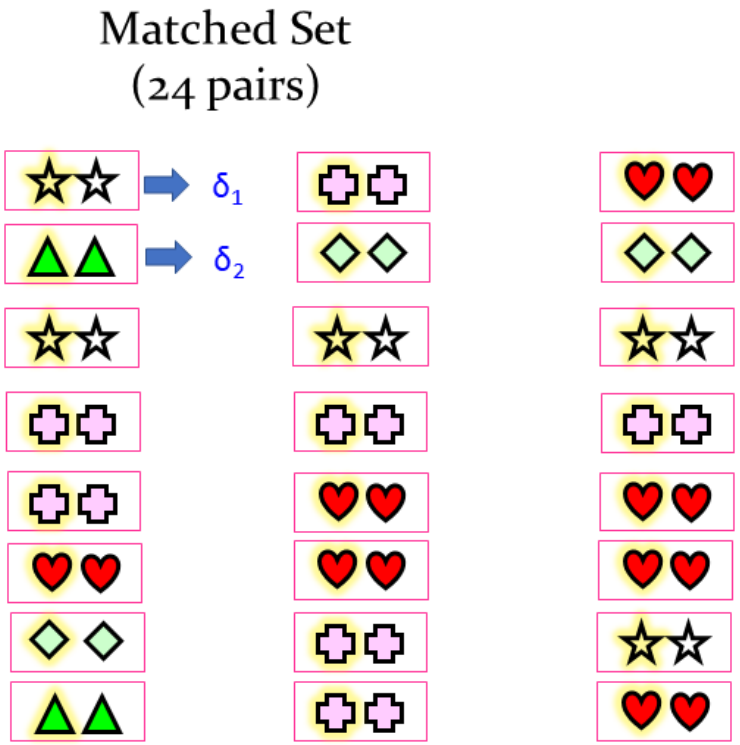
Characterize by propensity to receive the exposure...



Are baseline characteristics in balance?



Propensity Score Matching (1:1)

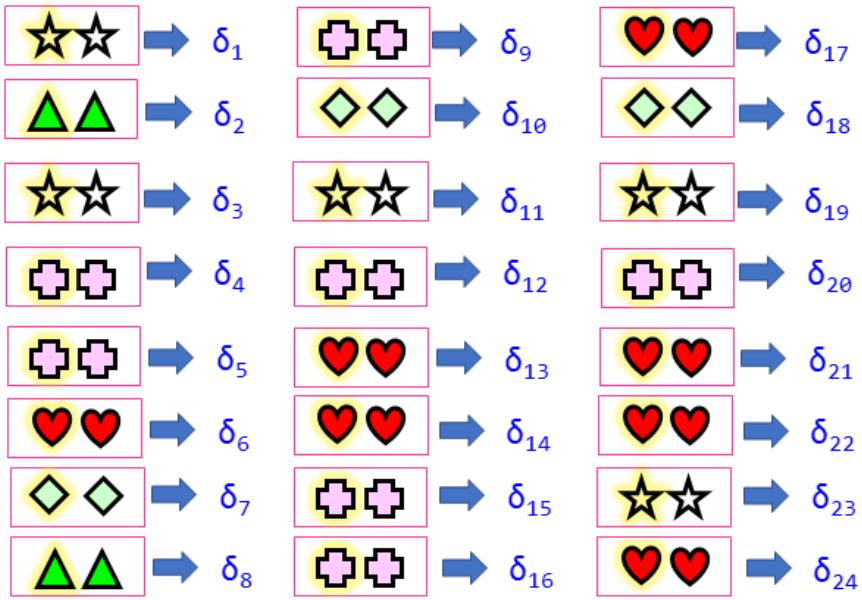


Within each matched pair, compare outcome in exposed subject to outcome in "not exposed" subject.

Estimated outcome effect Within a specific pair *j* is estimated by δ_j

Propensity Score Matching (1:1)

Matched Set
(24 pairs)



Within each matched pair, compare outcome in exposed subject to outcome in "not exposed" subject.

Use standard methods for matched samples (e.g., paired t tests) to estimate the causal effect of the exposure on the outcome based on the δ estimates from the pairs

What are we estimating here? ATT vs. ATE

Suppose we have an outcome Y, with potential outcomes Y(treated) and Y(control) and a treatment indicator Z, where Z = 1 if treated.

We can estimate the causal effect of Z on Y, using either an ATT or ATE approach.

- The average treatment effect on the treated (**ATT**) = E[Y(treated) - Y(control) | Z = 1], is the expected gain in outcome due to treatment for the population of people who were actually treated.
  - Most of the time, the ATT is the estimand we focus on in propensity score matching where we match one or more control patients (from a pool of such patients) to each treated patient.
  - The idea is to match the treated population closely.
- The average treatment effect (**ATE**) = E[Y(treated) - Y(control)], is the expected gain in outcome due to treatment for a randomly selected member of the entire population of interest.
  - The ATE estimate focuses on the population as a whole (treated + control).

Section 3

Multivariate Matching with the Propensity Score (the Aspirin Example)

Multivariate Matching with the Propensity Score

Match subjects so that they balance on multiple covariates using one scalar score[3].

- Goal: Emulate a RCT in matching, then use standard analyses to compare matched sets.
- Design: Treated subjects matched to people who didn't receive treatment but who had similar propensity to receive treatment (match the treated to untreated "clones.")

Multivariate Matching Mechanics

- Close but inexact PS matching on a large pool of covariates removes most of the bias due to those covariates
  - Assessing the Quality of the Matching
  - Checking for Covariate Balance
- Key Example: Aspirin Use and Mortality (Gum, 2001)

[3] Seminal papers: Rosenbaum and Rubin (1985, 1983, 1984)

Aspirin Use and All-Cause Mortality Among Patients Being Evaluated for Known or Suspected Coronary Artery Disease

A Propensity Analysis

Patricia A. Gum, MD
Maran Thamilarasan, MD
Junko Watanabe, MD
Eugene H. Blackstone, MD
Michael S. Lauer, MD

Context Although aspirin has been shown to reduce cardiovascular morbidity and short-term mortality following acute myocardial infarction, the association between its use and long-term all-cause mortality has not been well defined.

Objectives To determine whether aspirin is associated with a mortality benefit in stable patients with known or suspected coronary disease and to identify patient characteristics that predict the maximum absolute mortality benefit from aspirin.

Aspirin Use and Mortality (Gum 2001)

6174 consecutive adults at CCF undergoing stress echocardiography for evaluation of known or suspected coronary disease[4].

- 2310 (37%) were taking aspirin (treatment).
- Main Outcome: all-cause mortality
  - Median follow-up: 3.1 years

Analysis without covariates:

- 4.5% of the aspirin and 4.5% of the non-aspirin patients died.
- The unadjusted hazard ratio was 1.08 (0.85, 1.39).

[4] Gum PA et al. 2001

Adjustment for Covariates in Gum (2001)

- Demographics (Age, Sex)
- Cardiovascular risk factors
- Coronary disease history
- Use of other medications
- Ejection fraction
- Exercise capacity
- Heart rate recovery
- Echocardiographic ischemia

Adjusting for all of those factors in a regression model, then aspirin use is now associated with reduced mortality.

- Hazard Ratio 0.67, with 95% CI (0.51, 0.87)

Gum (2001) Table 1

Table 1. Baseline and Exercise Characteristics According to Aspirin Use*

Variable	Aspirin (n = 2310)	No Aspirin (n = 3864)	*P* Value	$\Delta_{A\text{-}No}$	Δ_{Std}
Demographics					
Age, mean (SD), y	62 (11)	56 (12)	<.001	6.0	**52.1**
Men, No. (%)	1779 (77)	2167 (56)	<.001	20.9	**45.5**
Clinical history					
Diabetes, No. (%)	388 (17)	432 (11)	<.001	5.6	**16.2**
Hypertension, No. (%)	1224 (53)	1569 (41)	<.001	12.4	**25.0**
Tobacco use, No. (%)	234 (10)	500 (13)	.001	-2.8	-8.8
Prior coronary artery disease, No. (%)	1609 (70)	778 (20)	<.001	49.5	**114.8**
Prior coronary artery bypass graft, No. (%)	689 (30)	240 (6)	<.001	23.6	**64.6**
Prior percutaneous coronary intervention, No. (%)	667 (29)	148 (4)	<.001	25.0	**72.0**
Prior Q-wave MI, No. (%)	369 (16)	285 (7)	<.001	8.6	**27.0**
Atrial fibrillation, No. (%)	27 (1)	55 (1)	.04	-0.3	-2.3
Congestive heart failure, No. (%)	127 (6)	178 (5)	.12	0.9	4.1
Medication use					
Digoxin use, No. (%)	171 (7)	216 (6)	.004	1.8	7.4
β-Blocker use, No. (%)	811 (35)	550 (14)	<.001	20.9	**49.9**
Diltiazem/verapamil use, No. (%)	452 (20)	405 (10)	<.001	9.1	**25.6**
Nifedipine use, No. (%)	261 (11)	283 (7)	<.001	4.0	**13.7**
Lipid-lowering therapy, No. (%)	775 (34)	380 (10)	<.001	23.7	**60.1**
ACE inhibitor use, No. (%)	349 (15)	441 (11)	<.001	3.7	**10.9**

Using Standardized Differences to Quantify Covariate Imbalance

For continuous variables,

$$\Delta_{Std} = \frac{100(\bar{x}_{ASA} - \bar{x}_{No})}{\sqrt{\frac{s^2_{ASA} + s^2_{No}}{2}}}$$

For binary variables,

$$\Delta_{Std} = \frac{100(p_{ASA} - p_{No})}{\sqrt{\frac{p_{ASA}(1-p_{ASA}) + p_{No}(1-p_{No})}{2}}}$$

Beta-Blocker	Aspirin	No Aspirin	Δ_{Std}
Before Match	35.1% (811/2310)	14.2% (550/3864)	49.9%
After Match	26.1% (352/1351)	26.5% (358/1351)	-1.0%

Gum (2001) Table 1 (continued)

Table 1. Baseline and Exercise Characteristics According to Aspirin Use*

Variable	Aspirin (n = 2310)	No Aspirin (n = 3864)	*P* Value	$\Delta_{A\text{-}No}$	Δ_{Std}
Cardiovascular assessment and exercise capacity					
Body mass index, mean (SD), kg/m^2	29 (5)	30 (7)	<.001	-1	**-16.4**
Ejection fraction, mean (SD), %	50 (9)	53 (7)	<.001	-3	**-37.2**
Resting heart rate, mean (SD), beats/min	74 (13)	79 (14)	<.001	-5	**-37.0**
Resting blood pressure, mean (SD), mm Hg					
Systolic	141 (21)	138 (20)	<.001	3	**14.6**
Diastolic	85 (11)	86 (11)	.04	-1	**-9.1**
Purpose of test to evaluate chest pain, No. (%)	300 (13)	468 (12)	.31	0.9	2.6
Mayo Risk Index ≥1, No. (%)†	2021 (87)	2517 (65)	<.001	22.3	**54.5**
Peak exercise capacity, mean (SD), METs					
Men	8.6 (2.4)	9.1 (2.6)	<.001	-0.5	**-20.0**
Women	6.6 (2.0)	7.3 (2.1)	<.001	-0.7	**-34.1**
Heart rate recovery, mean (SD), beats/min	28 (11)	30 (12)	<.001	-2.0	**-17.4**
Ischemic ECG changes with stress, No. (%)	430 (24)	457 (14)	<.001	6.8	**19.0**
Echocardiographic left ventricular ejection fraction ≤40%, No. (%)	321 (14)	226 (6)	<.001	8.0	**27.2**
Stress-induced ischemia on echocardiography, No. (%)	495 (21)	436 (11)	<.001	10.1	**27.7**
Fair or poor physical fitness for age and sex,[13] No. (%)	714 (31)	1248 (38)	.26	-1.4	-3.0

*MI indicates myocardial infarction; ACE, angiotensin-converting enzyme; MET, metabolic equivalent task; and ECG, electrocardiogram.
†The Mayo Risk Index is described in the "Methods" section.

Pre-Matching Characteristics by Aspirin Use

Do the aspirin and non-aspirin groups show important differences in distribution at baseline?

- At baseline, aspirin patients display higher risk of mortality, in general
  - they are older, more likely to be male, and more likely to have a clinical history
  - they are more likely to be on other medications than non-aspirin subjects
  - their cardiovascular assessments are (generally) worse and have worse exercise capacity
- The table reports on 31 characteristics prior to matching
  - 24 of 31 have p values below 0.001, one more is p = 0.001, and two more are p = 0.04
  - 25 of 31 have standardized differences of more than 10%, and six are more than 50%

Propensity Score Matching

For each patient, we have a propensity score.

1. Randomly select an Aspirin user.
2. Match to the non-user with closest propensity score (within some limit or matching within "calipers")
3. Eliminate both patients from pool, and repeat until you cannot find an acceptable match.

- Could match a non-user with Propensity Score inside "calipers" who matches exactly on characteristic X,
- Match non-user with Propensity score inside "calipers" and smallest "distance" on some pre-specified covariates.

Matching on Gender within PS Calipers

1. Shuffle "treatment" patients, and select one.
2. Find all "non-treated" with PS inside calipers (here we'll set calipers at treated PS $\pm$ .03).
3. Match patient within calipers of same gender.
4. Repeat until no more matches are possible.

Patient	Exposure	PS	Gender
A	Treated	.76	Male
B	Not Treated	.77	Female
C	Not Treated	.74	Male
D	Not Treated	.80	Male

Gum (2001) Matching Approach (Greedy and Incomplete):

- Tried to match each aspirin user to a unique non-user with a propensity score that was identical to five digits.
- If not possible, proceeded to a 4-digit match, then 3-digit, 2-digit, and finally a 1-digit match (i.e., propensity scores within .099).
- **Result**: matches for 1,351 (58%) of the 2,310 aspirin patients to 1,351 unique non-users.

Baseline Characteristics According to Aspirin Use (after matching)

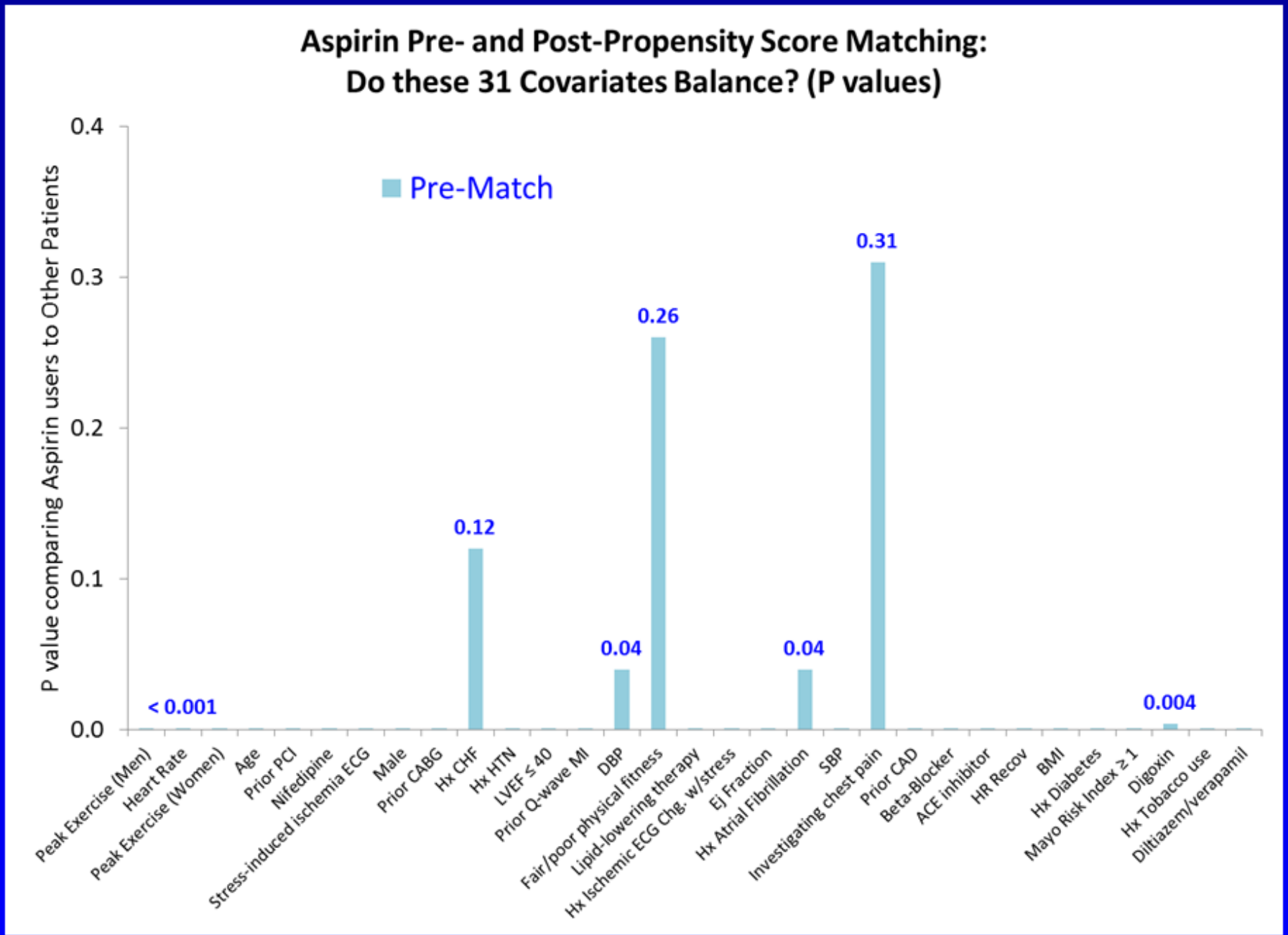
Variable	Aspirin* (n = 1351)	No Aspirin* (n = 1351)	P value
Age, years	60 (11)	61 (11)	.16
Body mass index, kg/m^2	29 (6)	29 (6)	.83
Ejection fraction, %	51 (8)	51 (9)	.65
Resting heart rate, beats/min	77 (13)	76 (14)	.13
Resting systolic BP, mm Hg	141 (21)	141 (21)	.68
Resting diastolic BP, mm Hg	85 (11)	86 (11)	.57
Heart rate recovery, beats/min	28 (12)	28 (11)	.82
Peak exercise cap., men (METs)	8.7 (2.5)	8.3 (2.5)	**.01**
Peak exercise capacity, women	6.5 (2.0)	6.7 (2.0)	.13

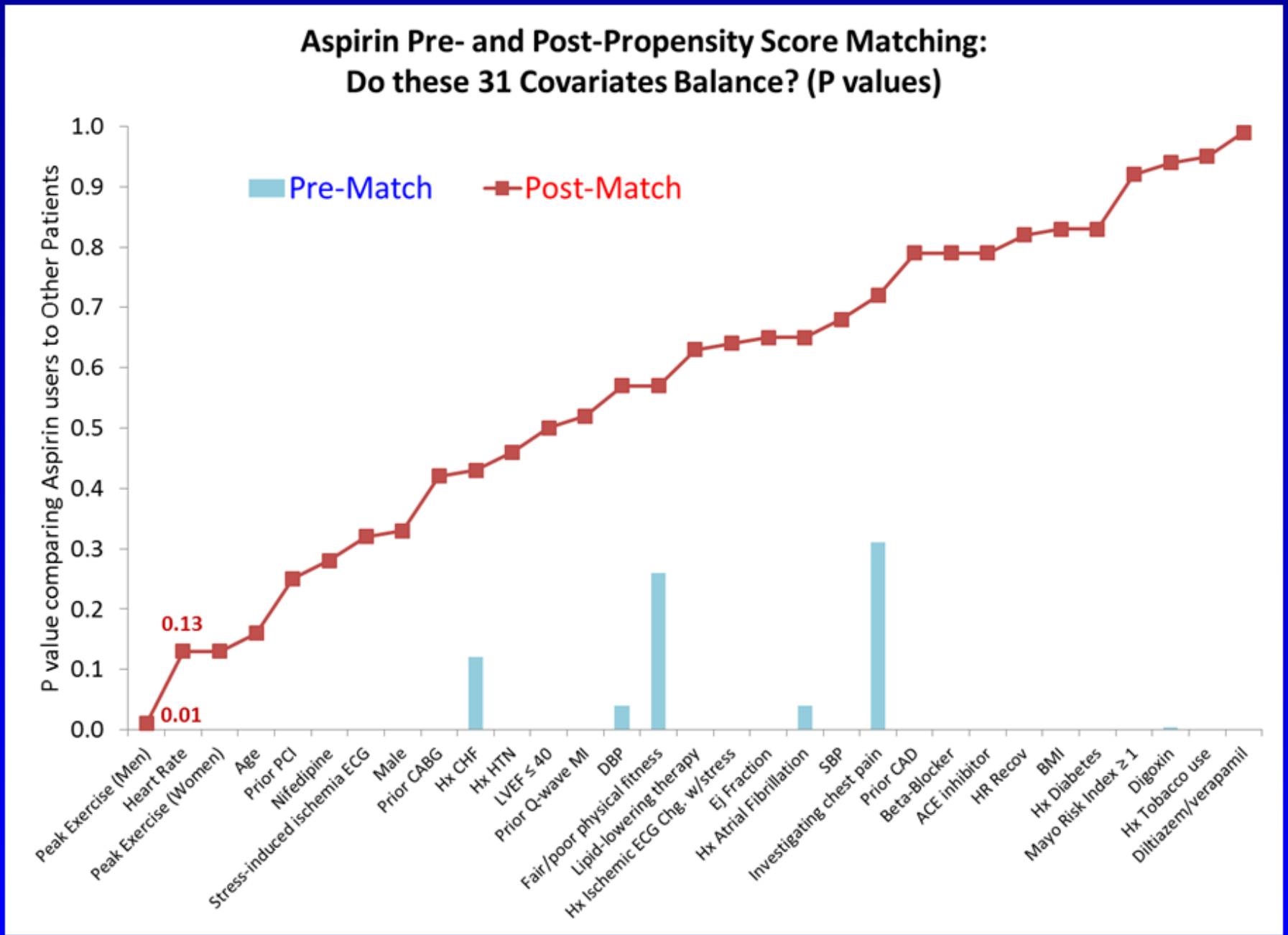
*Cells contain mean (SD)

Baseline Characteristics by Aspirin Use [%] (after matching)

Variable	Aspirin (n = 1351)	No Aspirin (n = 1351)	P value
Men	70.4	72.1	.33
Clinical history: diabetes	15.0	15.3	.83
hypertension	50.3	51.7	.46
prior coronary artery disease	48.3	48.8	.79
congestive heart failure	5.8	6.6	.43
Medication use: Beta-blocker	26.1	26.5	.79
ACE inhibitor	15.5	15.8	.79

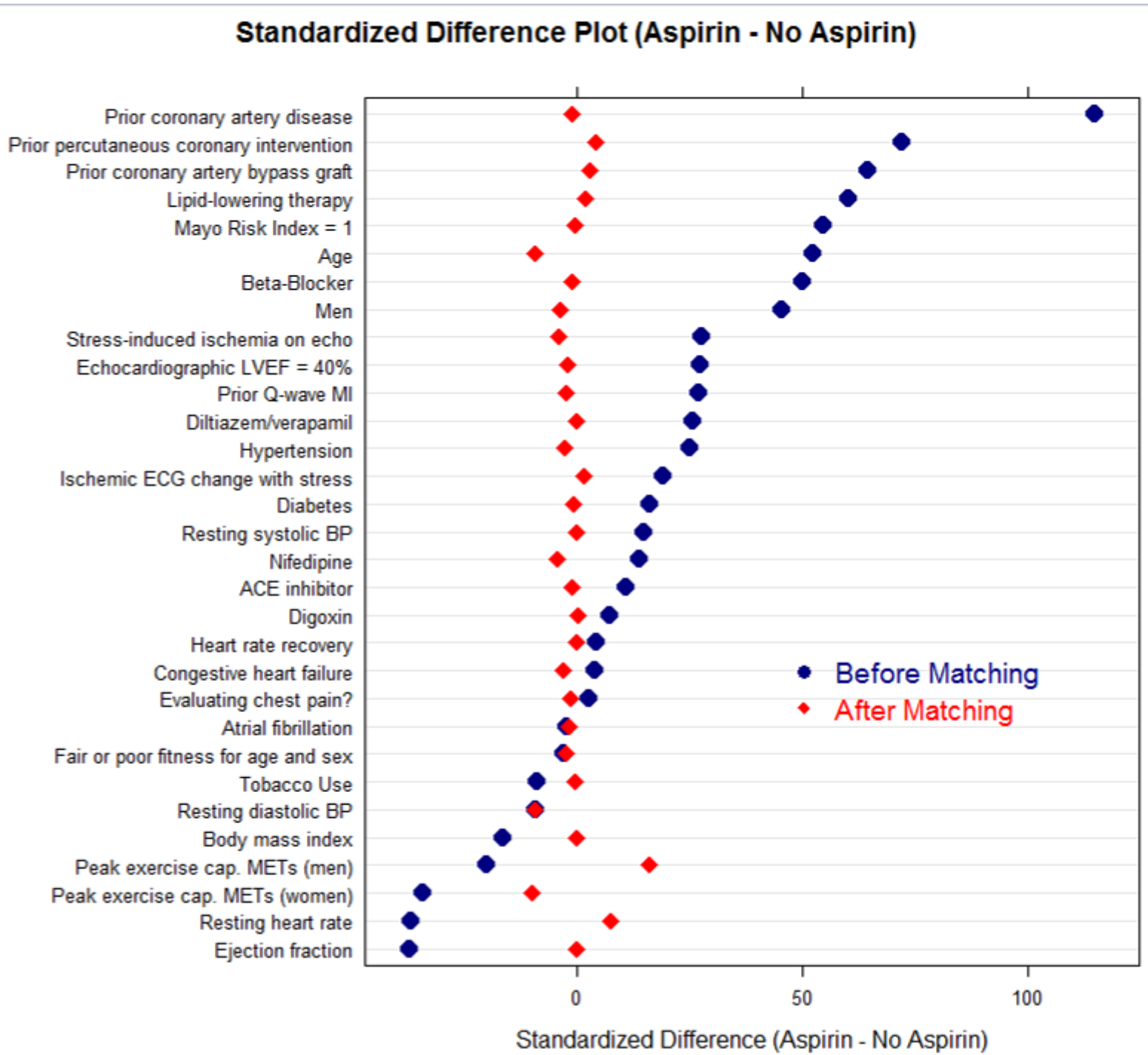
- Baseline characteristics similar in matched users and non-users.
- 30 of 31 covariates show NS difference between matched users and non-users. [Peak exercise capacity for men is p = .01]


Aspirin Pre- and Post-Propensity Score Matching:
Do these 31 Covariates Balance? (P values)
P value comparing Aspirin users to Other Patients
0.4
0.3
0.2
0.1
0.0
Pre-Match
< 0.001
0.12
0.04
0.26
0.04
0.31
0.004
Peak Exercise (Men)
Heart Rate
Peak Exercise (Women)
Age
Prior PCI
Nifedipine
Stress-induced ischemia ECG
Male
Prior CABG
Hx CHF
Hx HTN
LVEF ≤ 40
Prior Q-wave MI
DBP
Fair/poor physical fitness
Lipid-lowering therapy
Hx Ischemic ECG Chg. w/stress
Ej Fraction
Hx Atrial Fibrillation
SBP
Investigating chest pain
Prior CAD
Beta-Blocker
ACE inhibitor
HR Recov
BMI
Hx Diabetes
Mayo Risk Index ≥ 1
Digoxin
Hx Tobacco use
Diltiazem/verapamil


Aspirin Pre- and Post-Propensity Score Matching:
Do these 31 Covariates Balance? (P values)
P value comparing Aspirin users to Other Patients
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Pre-Match
Post-Match
0.13
0.01
Peak Exercise (Men)
Heart Rate
Peak Exercise (Women)
Age
Prior PCI
Nifedipine
Stress-induced ischemia ECG
Male
Prior CABG
Hx CHF
Hx HTN
LVEF ≤ 40
Prior Q-wave MI
DBP
Fair/poor physical fitness
Lipid-lowering therapy
Hx Ischemic ECG Chg. w/stress
Ej Fraction
Hx Atrial Fibrillation
SBP
Investigating chest pain
Prior CAD
Beta-Blocker
ACE inhibitor
HR Recov
BMI
Hx Diabetes
Mayo Risk Index ≥ 1
Digoxin
Hx Tobacco use
Diltiazem/verapamil

Section 4

Standardized Difference Plot (Love Plot)

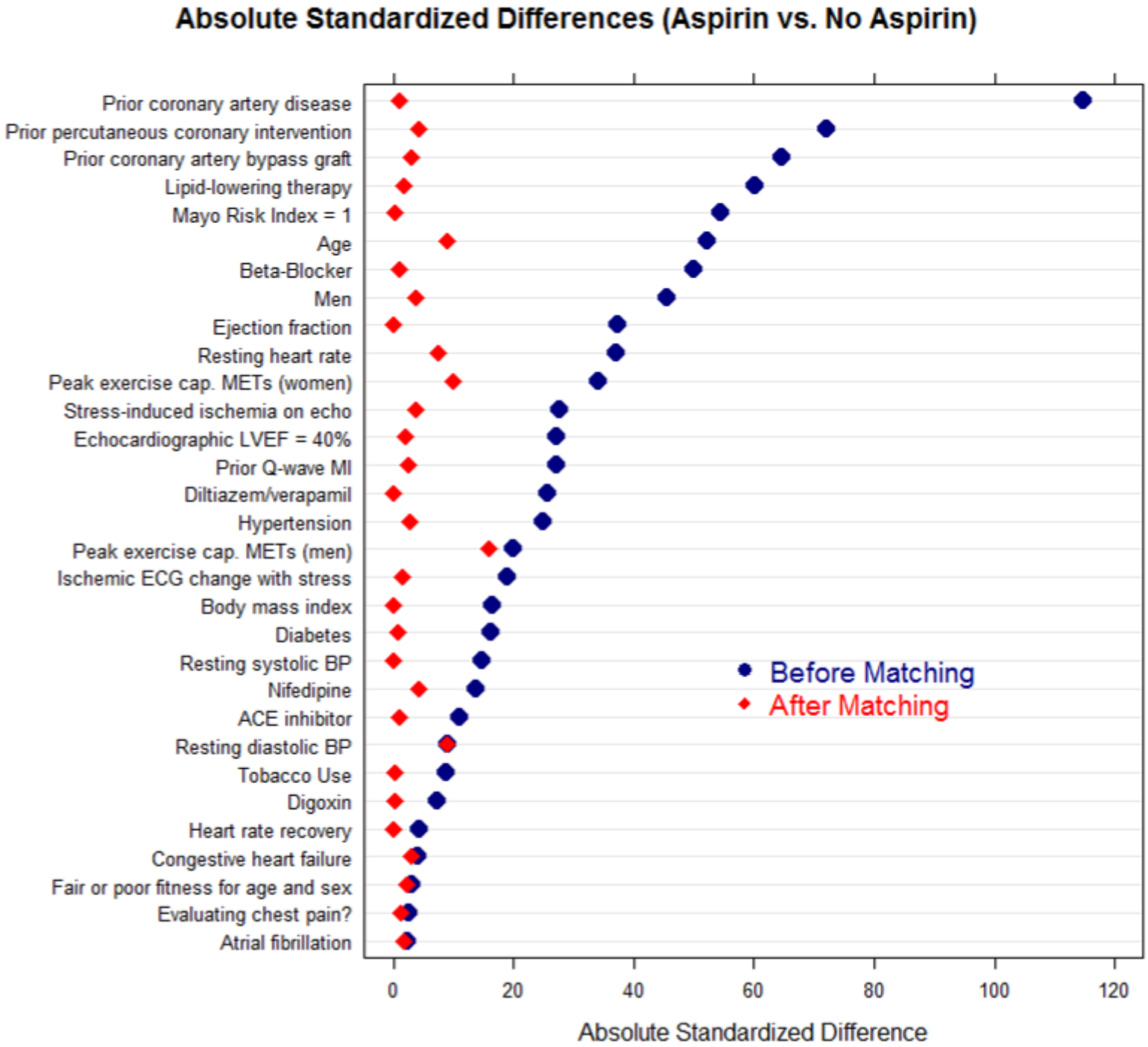

Standardized Difference Plot (Aspirin - No Aspirin)
Prior coronary artery disease
Prior percutaneous coronary intervention
Prior coronary artery bypass graft
Lipid-lowering therapy
Mayo Risk Index = 1
Age
Beta-Blocker
Men
Stress-induced ischemia on echo
Echocardiographic LVEF = 40%
Prior Q-wave MI
Diltiazem/verapamil
Hypertension
Ischemic ECG change with stress
Diabetes
Resting systolic BP
Nifedipine
ACE inhibitor
Digoxin
Heart rate recovery
Congestive heart failure
Evaluating chest pain?
Atrial fibrillation
Fair or poor fitness for age and sex
Tobacco Use
Resting diastolic BP
Body mass index
Peak exercise cap. METs (men)
Peak exercise cap. METs (women)
Resting heart rate
Ejection fraction
Before Matching
After Matching
0
50
100
Standardized Difference (Aspirin - No Aspirin)

Dotplots of Standardized Differences (Love Plots)

Why use them?

- Can work in a report or in Powerpoint, and in black and white or color.
- Has "at a glance" value, and doesn't require much "getting up to speed."
- Does not misstate the deviations.
- Follows general rules of good display (Tufte, Cleveland), i.e. good data-ink ratio, etc.
- "A-ha!" value. The plot helps the argument that the PS matching works when it does, and makes it clear where it doesn't when it doesn't.

We could also consider an **Absolute** Standardized Differences Plot (next slide)


Absolute Standardized Differences (Aspirin vs. No Aspirin)
Prior coronary artery disease
Prior percutaneous coronary intervention
Prior coronary artery bypass graft
Lipid-lowering therapy
Mayo Risk Index = 1
Age
Beta-Blocker
Men
Ejection fraction
Resting heart rate
Peak exercise cap. METs (women)
Stress-induced ischemia on echo
Echocardiographic LVEF = 40%
Prior Q-wave MI
Diltiazem/verapamil
Hypertension
Peak exercise cap. METs (men)
Ischemic ECG change with stress
Body mass index
Diabetes
Resting systolic BP
Nifedipine
ACE inhibitor
Resting diastolic BP
Tobacco Use
Digoxin
Heart rate recovery
Congestive heart failure
Fair or poor fitness for age and sex
Evaluating chest pain?
Atrial fibrillation
Before Matching
After Matching
0
20
40
60
80
100
120
Absolute Standardized Difference

What Should You Do About Residual Covariate Imbalance?

- Suppose a covariate appears seriously imbalanced after propensity matching.
  - Could make a regression adjustment for that covariate after matching.
  - Could use an additional or alternative measurement of the concept described by the covariate in the PS model.
  - Consider re-matching starting with a different random order of treated patients, or by a different standard.
  - Consider Mahalanobis distance matching within propensity score calipers.

Incomplete vs. Inexact Matching

- Trade-off between
  - Failing to match all treated subjects (incomplete)
  - Matching dissimilar subjects (inexact matching)
- Severe bias due to incomplete matching: so that it's usually better to match all treated subjects, then follow with analytical adjustments for residual imbalances in the covariates.
- But in practice (at least in the clinical literature), a bigger concern has been inexactness.
  - Certainly worthwhile to define the comparison group and carefully explore why subjects match.

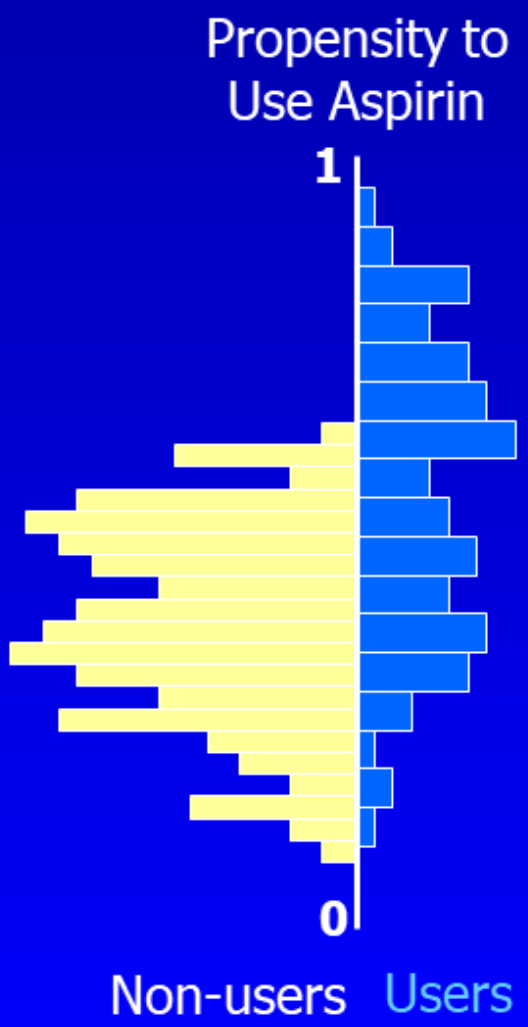
Which Aspirin Users Get Matched?

Generally, characteristics of unmatched aspirin users tend to indicate high propensity scores (to receive aspirin).

- Overall, 37% of patients were taking aspirin.
- The rate was much higher in some populations...
  - 67% of Prior CAD patients were taking aspirin.
  - So, prior CAD pts had higher propensity for aspirin.
  - 99.8% of unmatched aspirin users had prior CAD.
- Likely that unmatched users tended towards larger propensity scores than matched users

Who's Getting Matched Here?

Where Do The Propensity Scores Overlap?



Caveat: This simulation depicts what often happens.

Aspirin users we won't be able to match effectively

Aspirin users for whom we might realistically find a good match in our pool of non-users

Which Aspirin Users Get Matched?

- 652 of the 1351 matched aspirin users had had prior coronary artery disease (48.3%).
- 957 of the 959 unmatched aspirin users had had prior coronary artery disease (99.8%).

Variable	% of Matched	% of Unmatched	Standardized Difference
Prior CAD	48.3	99.8	-145
Prior PCI	12.3	52.2	-95
Lipid-low th.	20.8	51.5	-68
Prior CABG	18.6	45.7	-61
β-blocker	26.1	47.9	-46
Tobacco	11.9	7.3	+16

Matching with Propensity Scores

1,351 aspirin subjects matched well to 1,351 unique non-aspirin subjects

- Big improvement in covariate balance
- Table 1 for matched group looks like an RCT
- Can analyze the resulting matched pairs with standard methods (stratified Cox models, etc.)

Matching still incomplete (lots of possible bias here) and this isn't the best algorithm for matching, either...

Estimating the Hazard Ratios

During follow-up, 153 (6%) of the 2,702 matched patients died.

- In the matched group, aspirin use was associated with a lower risk of death (4% vs. 8%, p = 0.002)

Approach	n	Est. HR	95% CI
Full sample, no adjustment	6174	1.08	(0.85, 1.39)
Full sample, no PS, adj. for all covariates	6174	0.67	(0.51, 0.87)
PS-matched sample	2702	0.53	(0.38, 0.74)
PS-matched, adj. for PS and all covariates	2702	0.56	(0.40, 0.78)

Our PS-matched approaches here yield ATT estimates.

Aspirin Conclusions / Caveats

- Subjects included in this study *may* be a more representative sample of real world patients than an RCT would provide.
  - On the other hand, they were getting cardiovascular care at the Cleveland Clinic.
  - And there are some inclusion and exclusion criteria here, too.
- PS matching still isn't randomization, we can only account here for the factors that were measured, and only as well as the instruments can measure them.
- There's no information here on aspirin dose, aspirin allergy, duration of treatment or medication adjustments.

Statistical Concerns

- This isn't the best way to match, certainly.
- There's no formal assessment of sensitivity to hidden bias.
- Looks like they avoided the issue of missing data.

Dealing with Missing Data

What if we have missing covariate values[5]?

- The pattern of missing covariates is easy to balance
  - Add a missingness indicator variable for all covariates with `NA`
  - Then fill in values for those cases in the original variable before estimating PS
- Matching on this augmented PS will tend to balance the observed covariates and the **pattern** of missingness, but yields no guarantee that the missing values themselves are actually balanced.

[5] For more on these issues, try D'Agostino 1998 and D'Agostino and Rubin 2000

When is Matching A Good Choice?

Certain covariates are more easily controlled through matching in the design than through analytical adjustments.

- Typically these are covariates that classify subjects into many small categories.
- If matching isn't used, some categories may wind up with treated subjects and no controls, or vice versa.

Cost is an important consideration.

- If some covariate information is readily available, but other data are difficult to obtain or expensive, matching becomes more attractive.
  - If data come with negligible costs, matching during the design is less attractive.
  - Why? Suppose some controls are so different (at baseline) from the treated subjects that they will be of little use. Matching may stop you from collecting data on such controls.

Matching Conclusions, 1

Matching is a fundamental part of the toolbox. For a book-length treatment, I recommend Rosenbaum 2010.

- Propensity scores facilitate matching on multiple covariates at once.
  - Matching is especially attractive when covariates classify subjects into many small categories.
- Matching on a multivariate distance within PS calipers often beats matching on the PS alone, especially if you can pre-specify pivotal covariates.
  - Matching within PS calipers followed by additional matching on key prognostic covariates is an effective method for both reducing bias and understanding the effects of specific covariates.
  - Matching on `logit(PS)` rather than on raw `PS` can often improve yield.

Matching Conclusions, 2

- If match is incomplete, it's especially useful to consider both matching and non-matching analyses
- Optimal matches, full matches, cardinality matches, genetic matches and other more sophisticated matching approaches can be fruitful.
- Matching can be especially attractive if data are costly - we can match on what we have first, and then collect new data only on the pre-matched subjects.

Section 5

Propensity Scores: Not Just For Matching

Methods for Using Propensity Scores

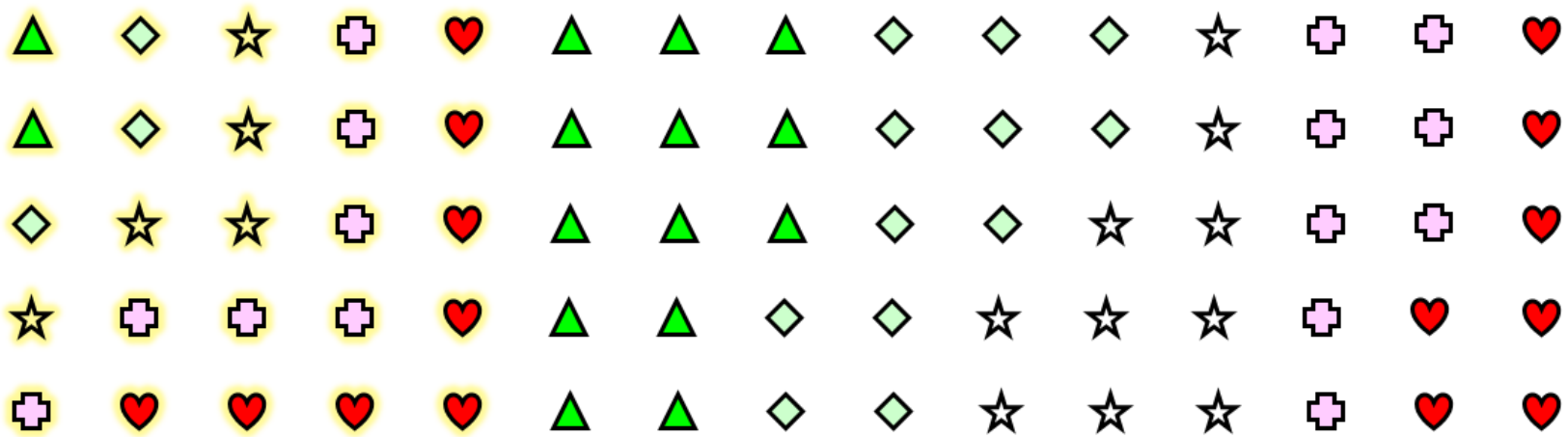
- Matching using the Propensity Score
- Direct (Regression) Adjustment using the Propensity Score
- Subclassification / Stratification on the Propensity Score
- Weighting using the Propensity Score
- Combining Approaches for More Robust Estimation

See Rosenbaum and Rubin 1983, 1984 and 1985 for foundational work.

Direct Adjustment for the Propensity Score

Model Without the Propensity Score

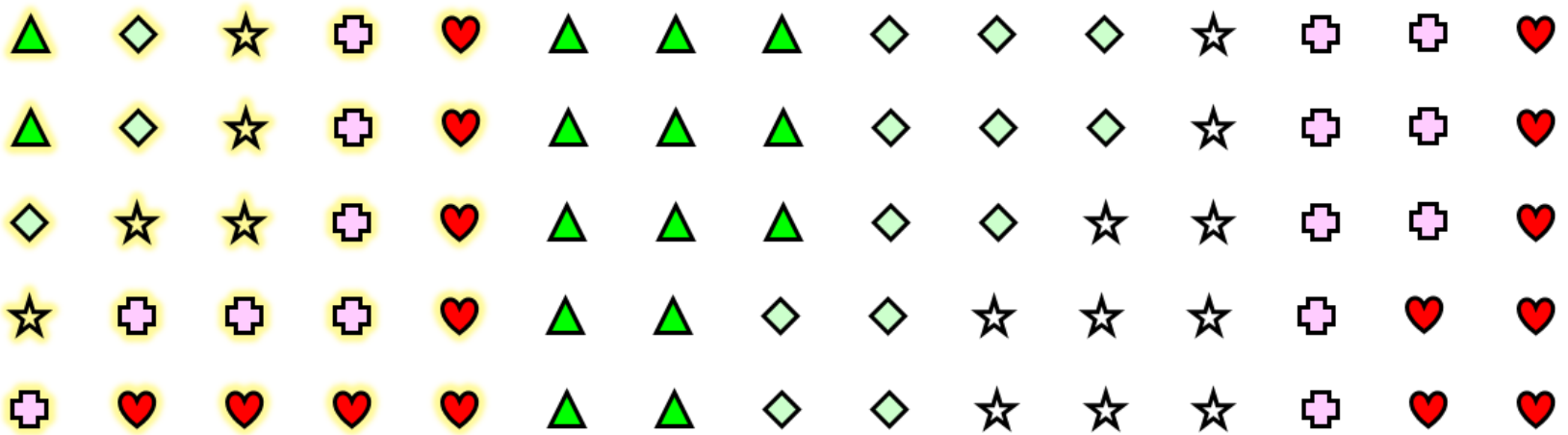
Outcome = β_0 + β_1*Exposure, for pool of 75 subjects



Direct Adjustment for Propensity Score

Outcome = $ß_0$ + $ß_1$*Exposure + $ß_2$*Propensity Score,

Again, across entire pool of 75 subjects



Direct Adjustment for the Propensity Score

Typically, we'll use the linear propensity score (the logit of the raw propensity score) here, to avoid problems with having propensity score estimates near 0 or 1.

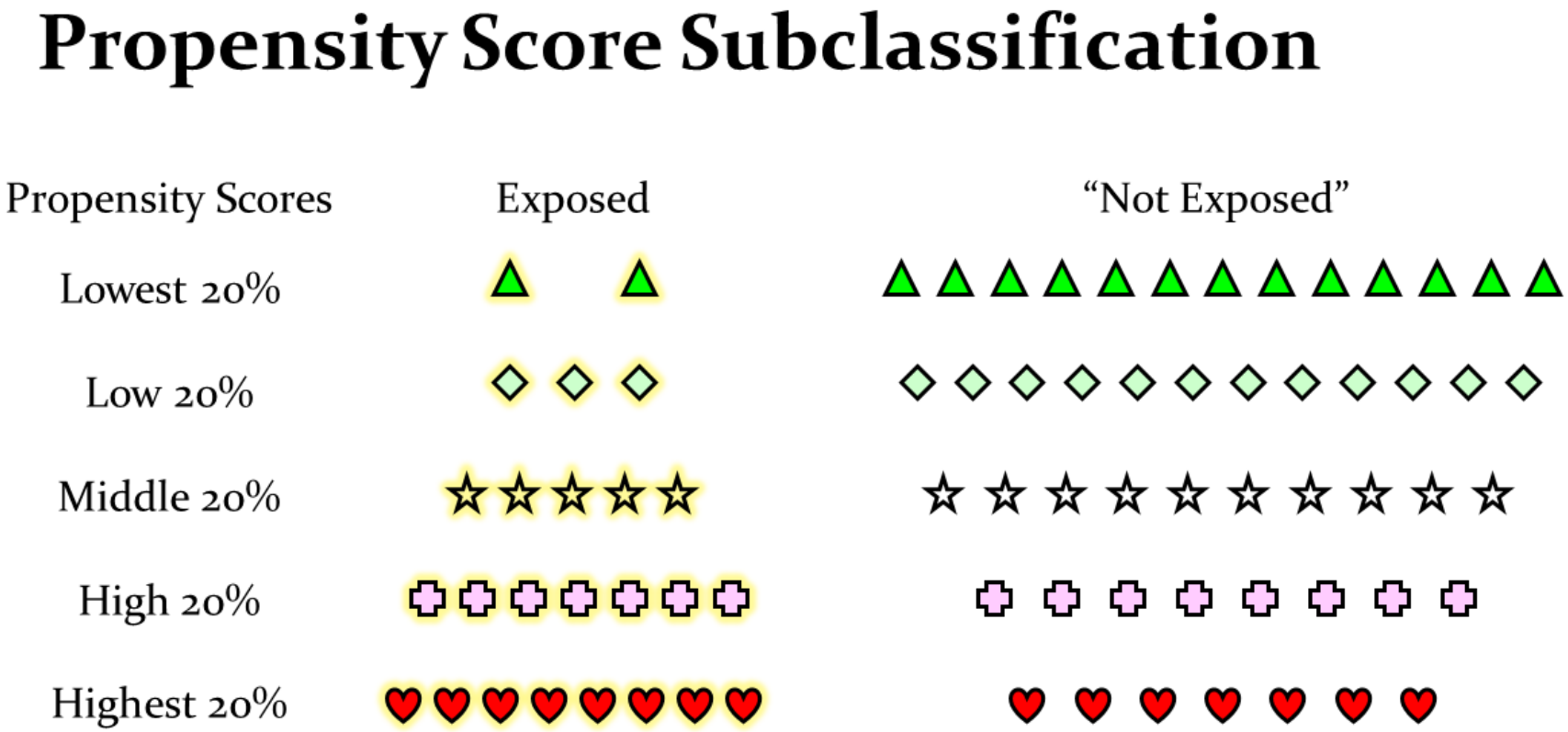
- The linear propensity score ranges across the real number line, rather than being restricted to 0 and 1.
- Our "Rubin's Rules" that help us think about the quality of balance necessary to justify regression models for our outcomes also work with linear propensity scores.
    - We'll discuss Rubin's Rules when we discuss Rubin 2001 later this semester.

Double Robust Estimates

Adjusting for the propensity score is often (if not usually) done in *combination* with other propensity score approaches, like matching or weighting to form what are called **double robust** estimates.

- For instance, we can match on the propensity score to obtain our matched sample, then further adjust in our outcome model using the (linear) propensity score again, or perhaps individual covariates that especially concern us with regard to our outcome of interest.
- We'll see that a similar approach is available to us with weighting + adjustment, or even subclassification + adjustment.

Propensity Score Subclassification / Stratification

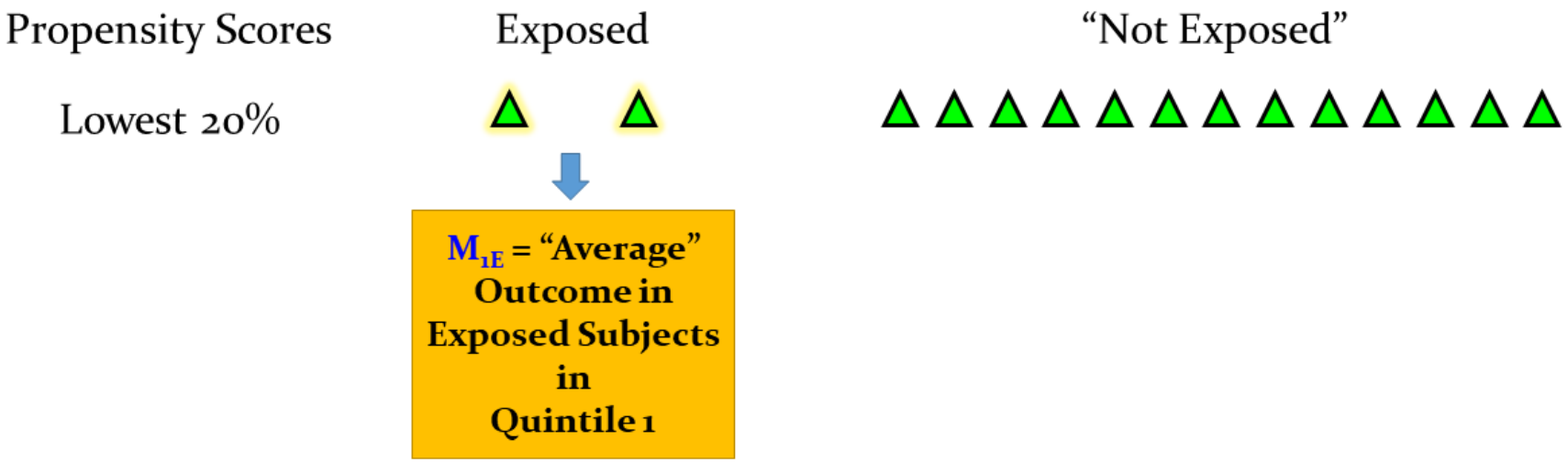


Propensity Score Subclassification

1. Split Subjects by Propensity Quintile

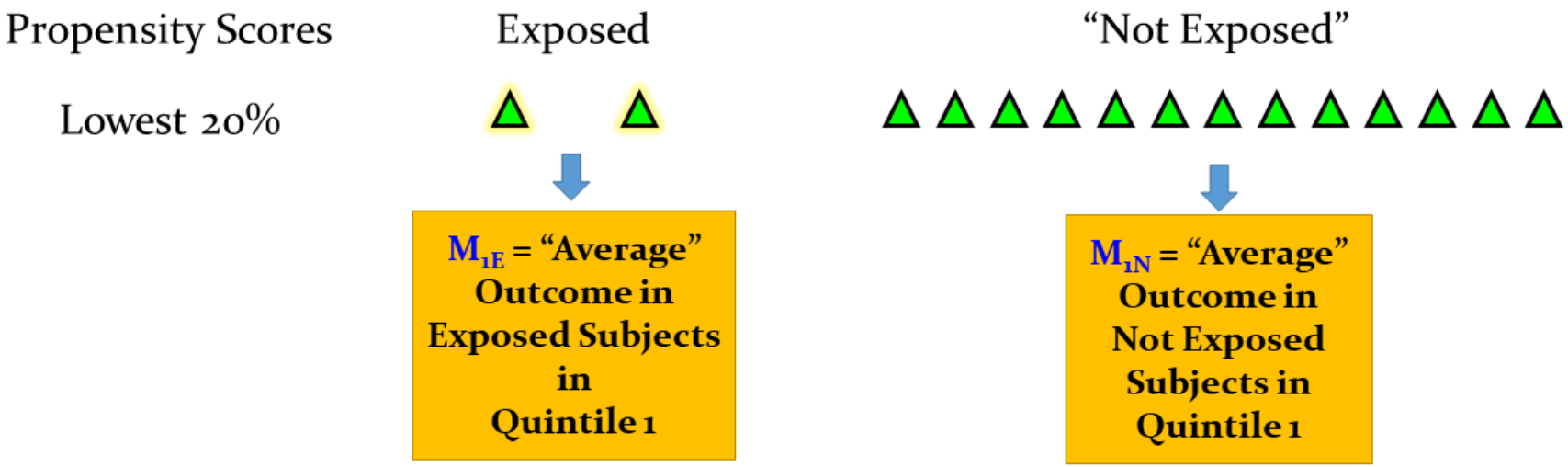
Propensity Score Subclassification

2. Estimate Effects Separately by Quintile



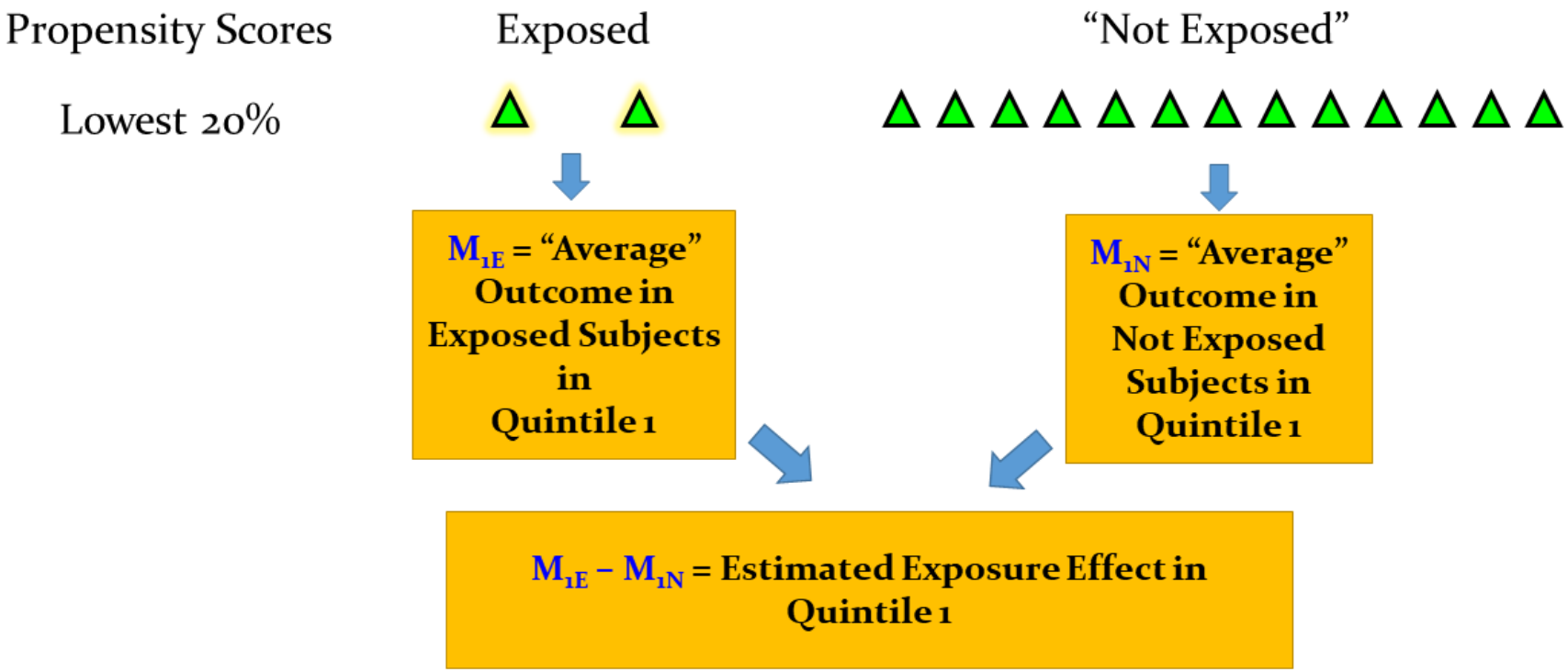
Propensity Score Subclassification

2. Estimate Effects Separately by Quintile



Propensity Score Subclassification

2. Estimate Effects Separately by Quintile



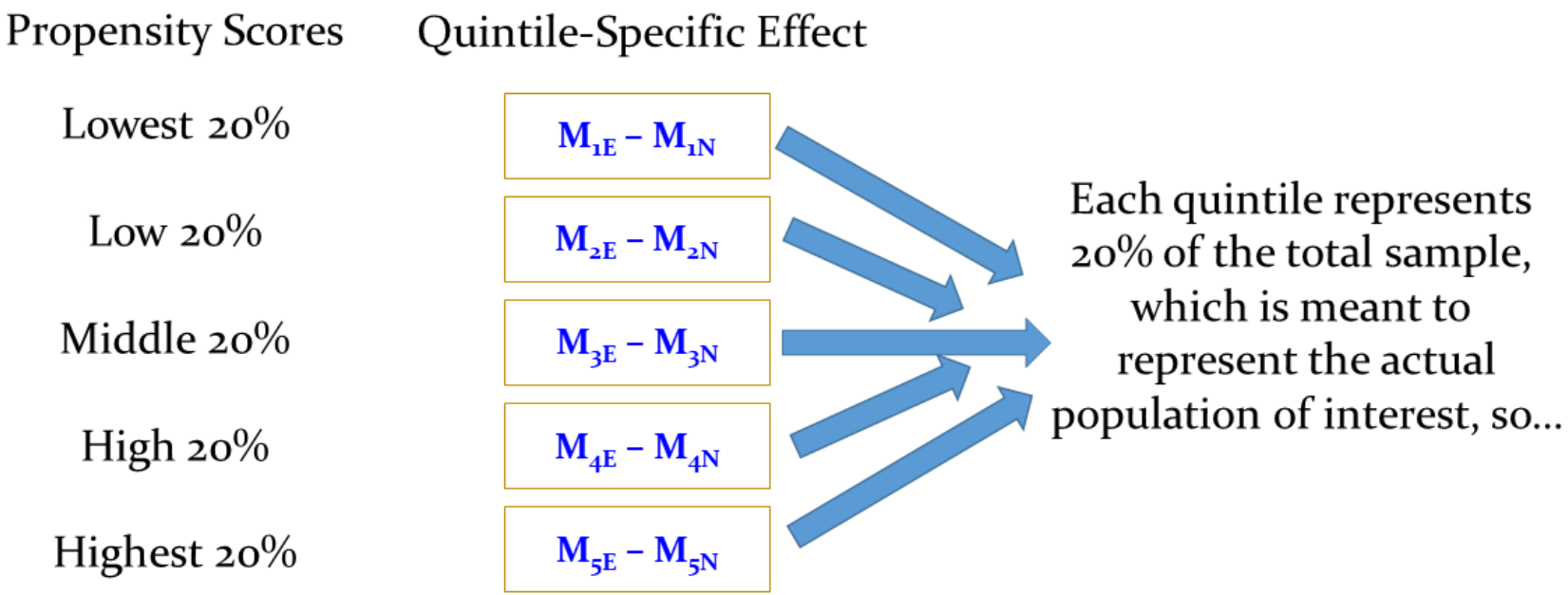
Propensity Score Subclassification

2. Estimate Effects Separately by Quintile

Propensity Scores	
Lowest 20%	$M_{1E} - M_{1N}$ = Estimated Exposure Effect in Quintile 1
Low 20%	$M_{2E} - M_{2N}$ = Estimated Exposure Effect in Quintile 2
Middle 20%	$M_{3E} - M_{3N}$ = Estimated Exposure Effect in Quintile 3
High 20%	$M_{4E} - M_{4N}$ = Estimated Exposure Effect in Quintile 4
Highest 20%	$M_{5E} - M_{5N}$ = Estimated Exposure Effect in Quintile 5

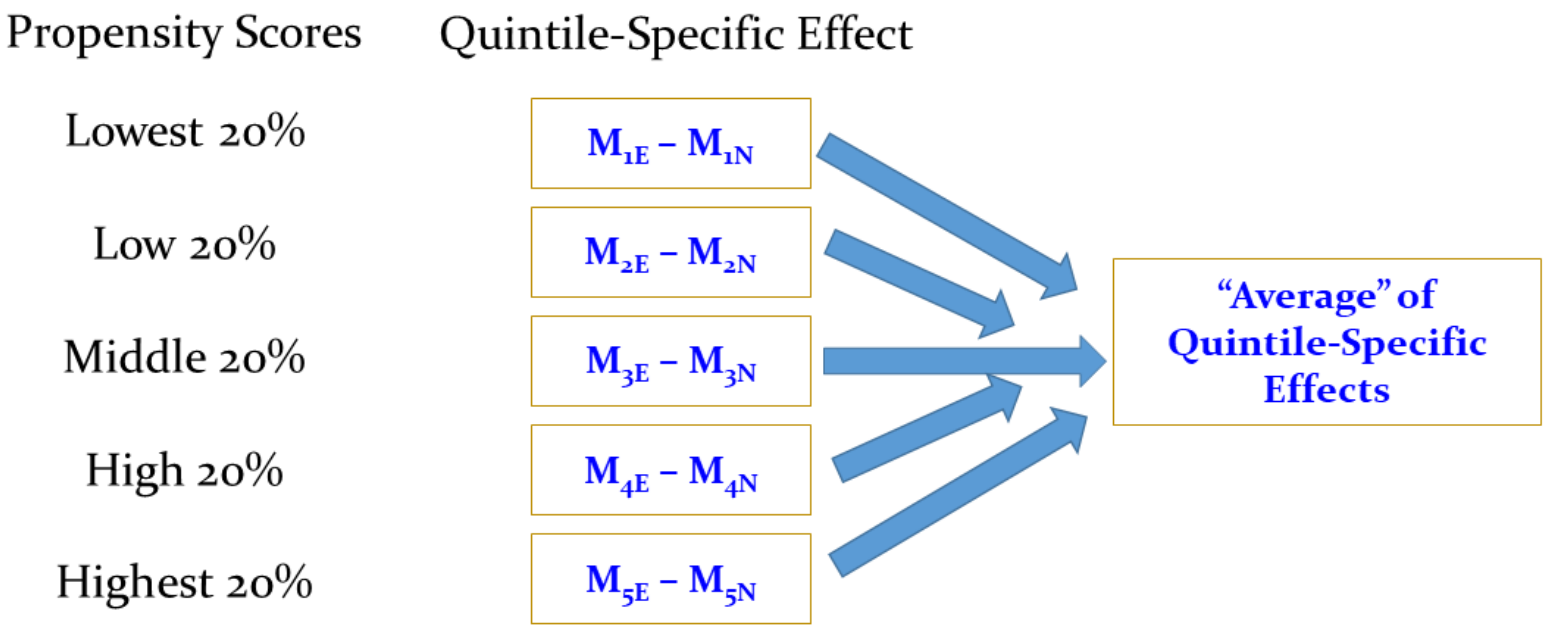
Propensity Score Subclassification

3. Combine Quintile-Specific Estimates



Propensity Score Subclassification

3. Combine Quintile-Specific Estimates



Propensity Score Weighting

Adjusting for the propensity score removes the bias associated with differences in the observed covariates in the exposed and control groups.

One way to implement this is to **reweight** exposed and control observations (or just controls, sometimes) to make them representative of the population of interest.

- PS methods generally lead to more reliable estimates of association than multiple regression, especially if there is a substantial selection or other overt bias.
- We can get the benefits of matching while still using all of the collected data.
- We can incorporate propensity weighting along with survey weighting, when oversampling is done, for instance.
- We can incorporate weighting with regression adjustment on the propensity score, producing a double robust estimate.

Propensity Score Weighting ("ATT")

All Exposed
get weight 1

Propensity Score Weighting ("ATT")

All Exposed
get weight 1

"Not Exposed"
unweighted

Propensity Score Weighting ("ATT")

All Exposed
get weight 1

"Not Exposed"
weighted

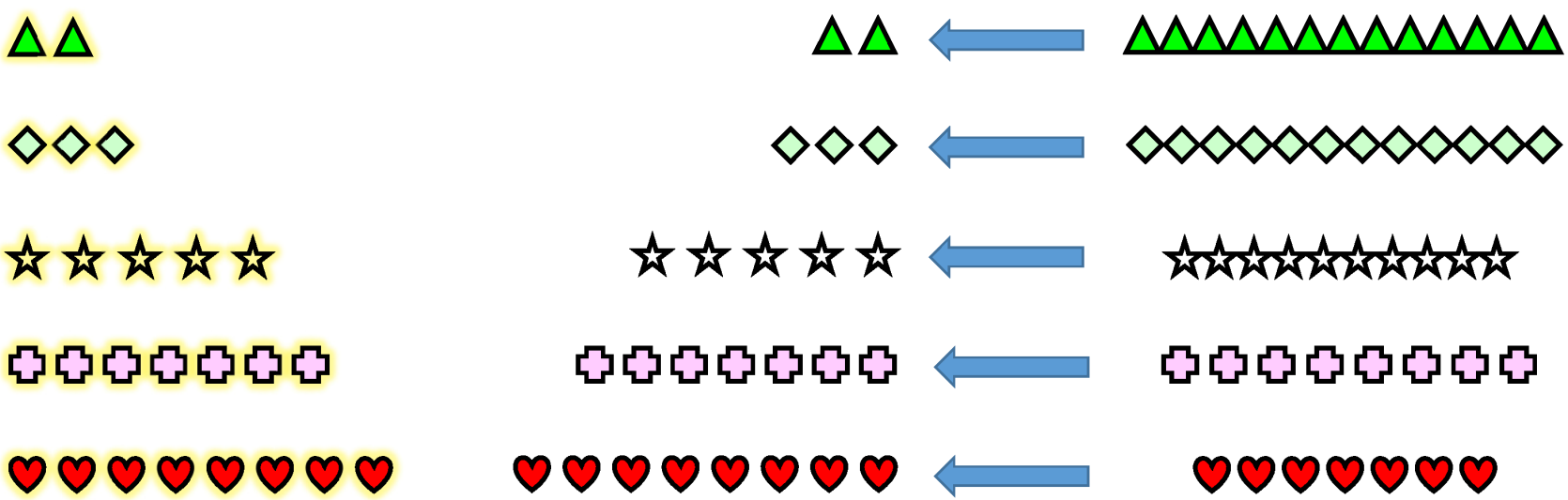
"Not Exposed"
unweighted

Propensity Score Weighting ("ATT")

All Exposed
get weight 1

"Not Exposed"
weighted

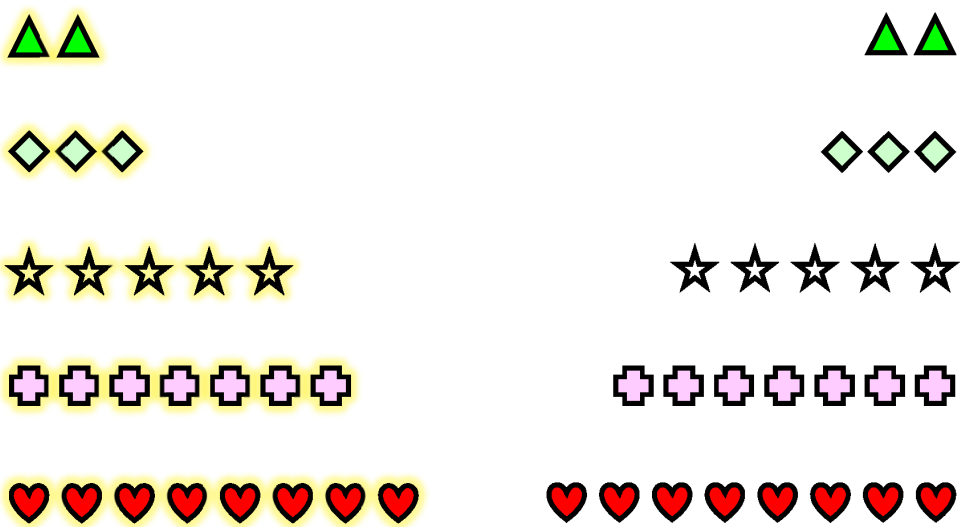
"Not Exposed"
unweighted



Propensity Score Weighting ("ATT")

All Exposed
get weight 1

"Not Exposed"
weighted



Average Outcome with Exposure

Outcome without Exposure (weighted)

"Weighted Average" Effect of Exposure on Outcome

ATT Weighting using the Propensity Score

ATT = average treatment effect on the treated

- Let every exposed (treated) subject's weight be 1.
- A control subject's weight is a function of its propensity for exposure

$$w_j = \frac{PS_j}{1 - PS_j}$$

ATT estimate = Average outcome for treated group - PS weighted outcome for control group

ATE Weighting using the Propensity Score

Alternatively, we can reweight both exposed and control patients to obtain an average treatment effect estimate[6].

- An exposed (treated) subject's weight is the inverse of its propensity score.

$$w_j = \frac{1}{PS_j}$$

- A control subject's weight is the inverse of one minus its propensity for exposure.

$$w_j = \frac{1}{1 - PS_j}$$

[6] For more, see Rubin 2001, and Lunceford and Davidian 2004

Section 6

The SUPPORT study

Studying Right Heart Catheterization in SUPPORT

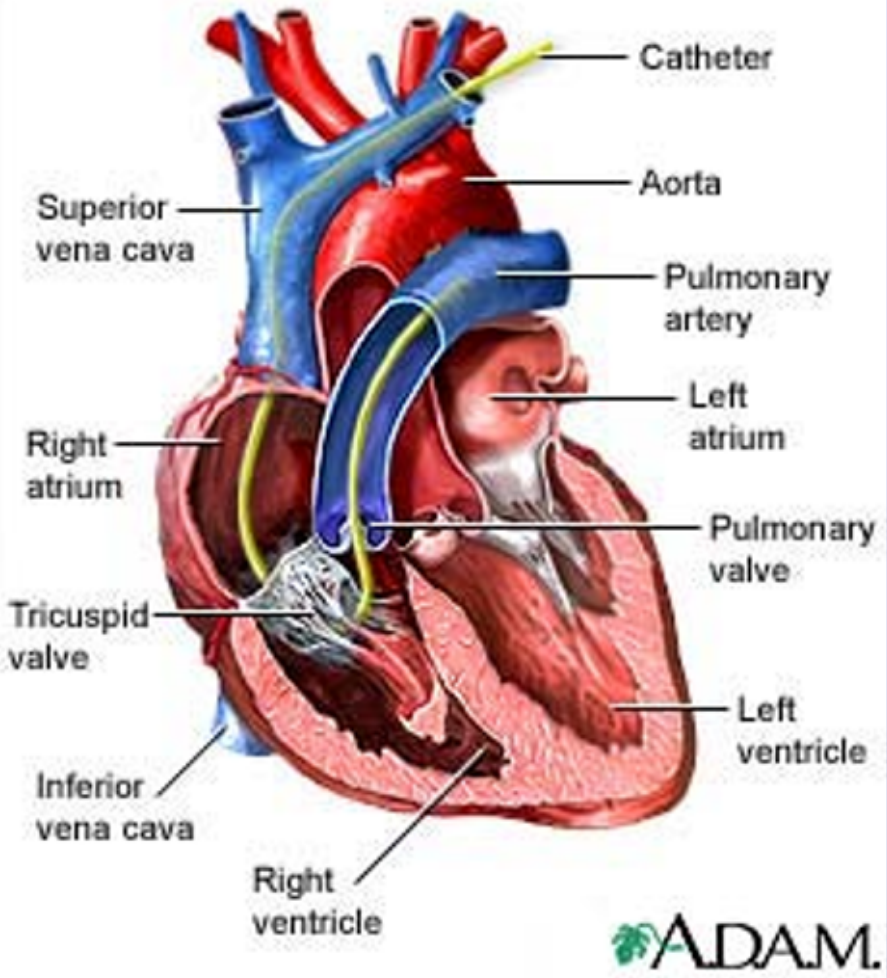
SUPPORT: Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments[7]

- Goal: Examine the association between the use of RHC during the first 24 hours in the ICU and outcomes
- Outcomes: survival, length of stay, intensity and costs of care
- Sample: 5,735 critically ill adult ICU patients in nine disease categories

Study was prospective!

[7] Connors et al. 1996

Right Heart / Swan-Ganz / Pulmonary Artery Catheterization



Pass a thin flexible tube:

1. **to obtain diagnostic information about the heart**
2. **for monitoring of heart function in the critically ill.**

http://www.nlm.nih.gov/medlineplus/ency/imagepages/18087.htm

Does the RHC do more harm than good?

Prior (small) observational studies comparing RHC to non-RHC patients:

- RR of death higher in RHC elderly patients than non-RHC elderly
- RR of death higher in RHC patients with acute MI than non-RHC patients with MI
- Patients with higher than expected RHC use had higher mortality

Big Problem: Selection Bias. Physicians (mostly) decide who gets RHC and who doesn't.

Why not a RCT?

- RHC directly measures cardiac function
- Some MDs believe RHC is necessary to guide therapy for some critically ill patients
- Procedure is very popular - existing studies haven't created equipoise

81 Characteristics used to predict PS(RHC usage)

- Age, Sex, Race
- Education, Income, Insurance
- Primary and Secondary Disease category
- Admission diagnosis category (12 levels)
- ADL and DASI 2 weeks before admission
- DNR status on day 1
- Cancer (none, local, metastasized)
- 2 month survival model
- Weight, temperature, BP, heart rate, respiratory rate
- Comorbid illness (13 categories)
- Body chemistry (pH, WBC, PaCO_2_, etc.)

Panel (7 specialists in clinical care) specified important variables related to the decision to use or not use a RHC.

RHC vs. Non-RHC patients

RHC patients were more likely to

- Be male, have private insurance, enter the study with ARF, MOSF or CHF

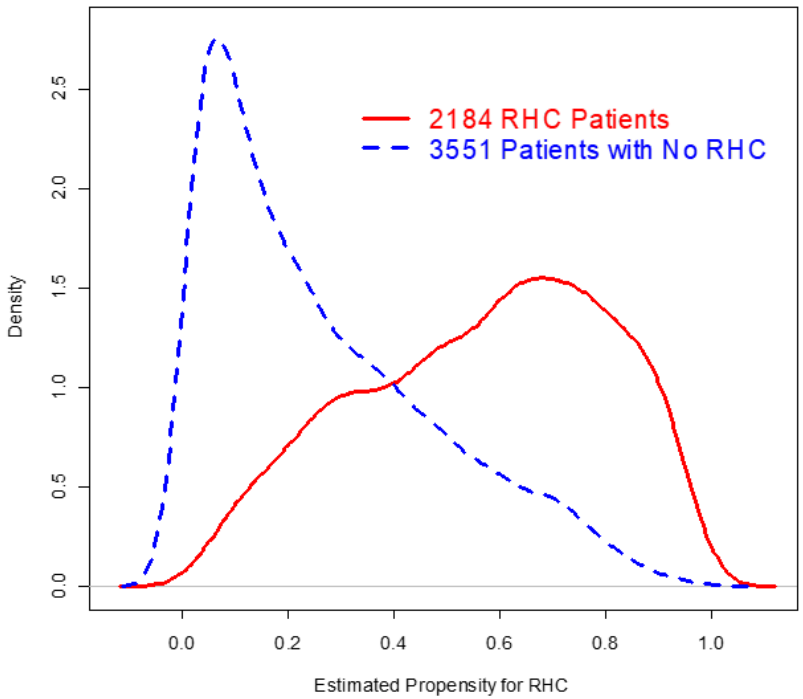
RHC patients were less likely to

- Be over 80 years old, have cancer, have a DNR order in the first 24 hours of hospitalization

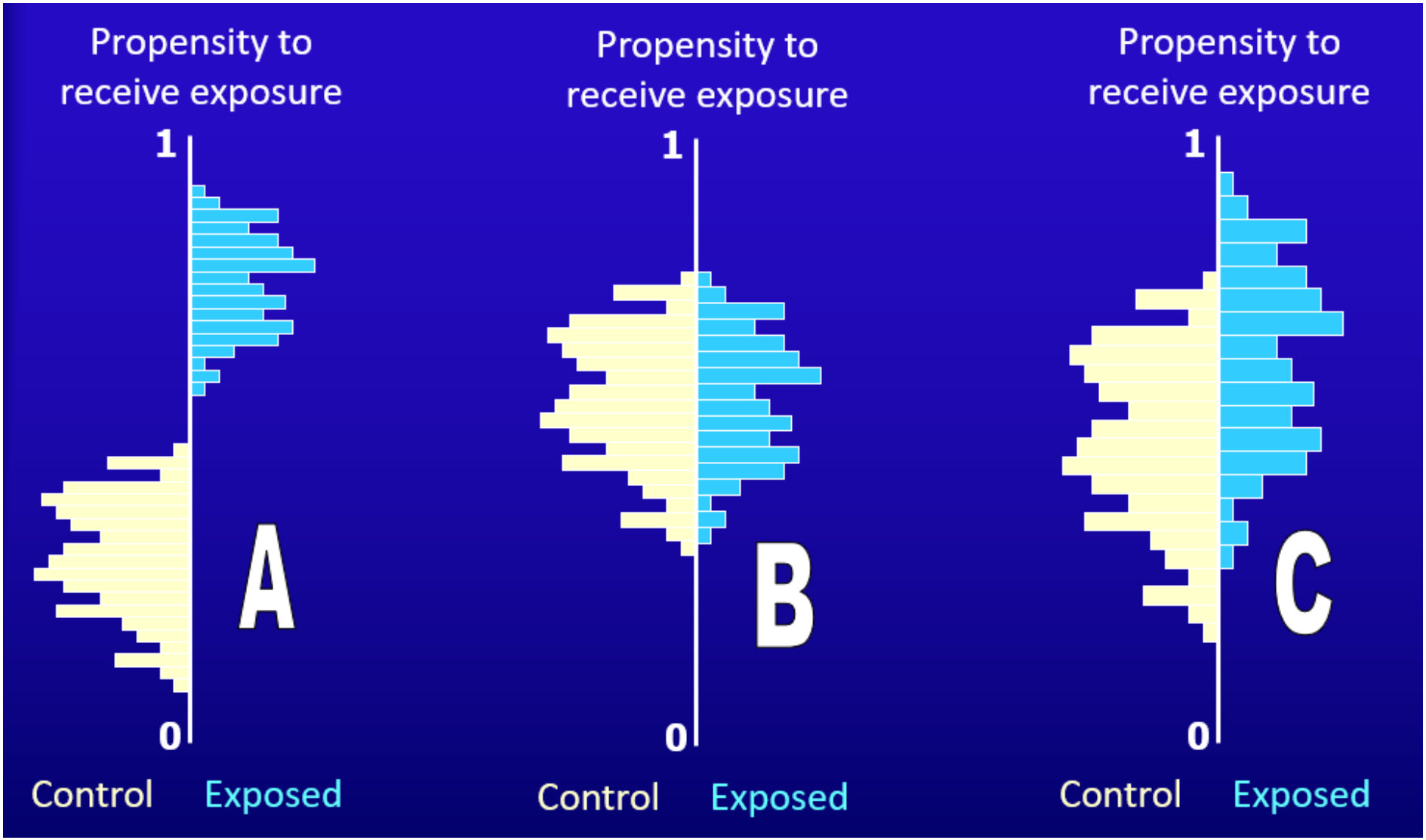
RHC patients had significantly

- Fewer comorbid conditions,
- More abnormal results of vital signs, WBC count, albumin, creatinine, etc.
- Lower model probability of 2-month survival

How Much Overlap do we see in the RHC data?



How Much Overlap do we want?



Right Heart Catheterization and the Perils of Selective Weighting

- 5,735 hospitalized patients in SUPPORT study
  - 2,184 treated (RHC) and 3,551 controls (no RHC).

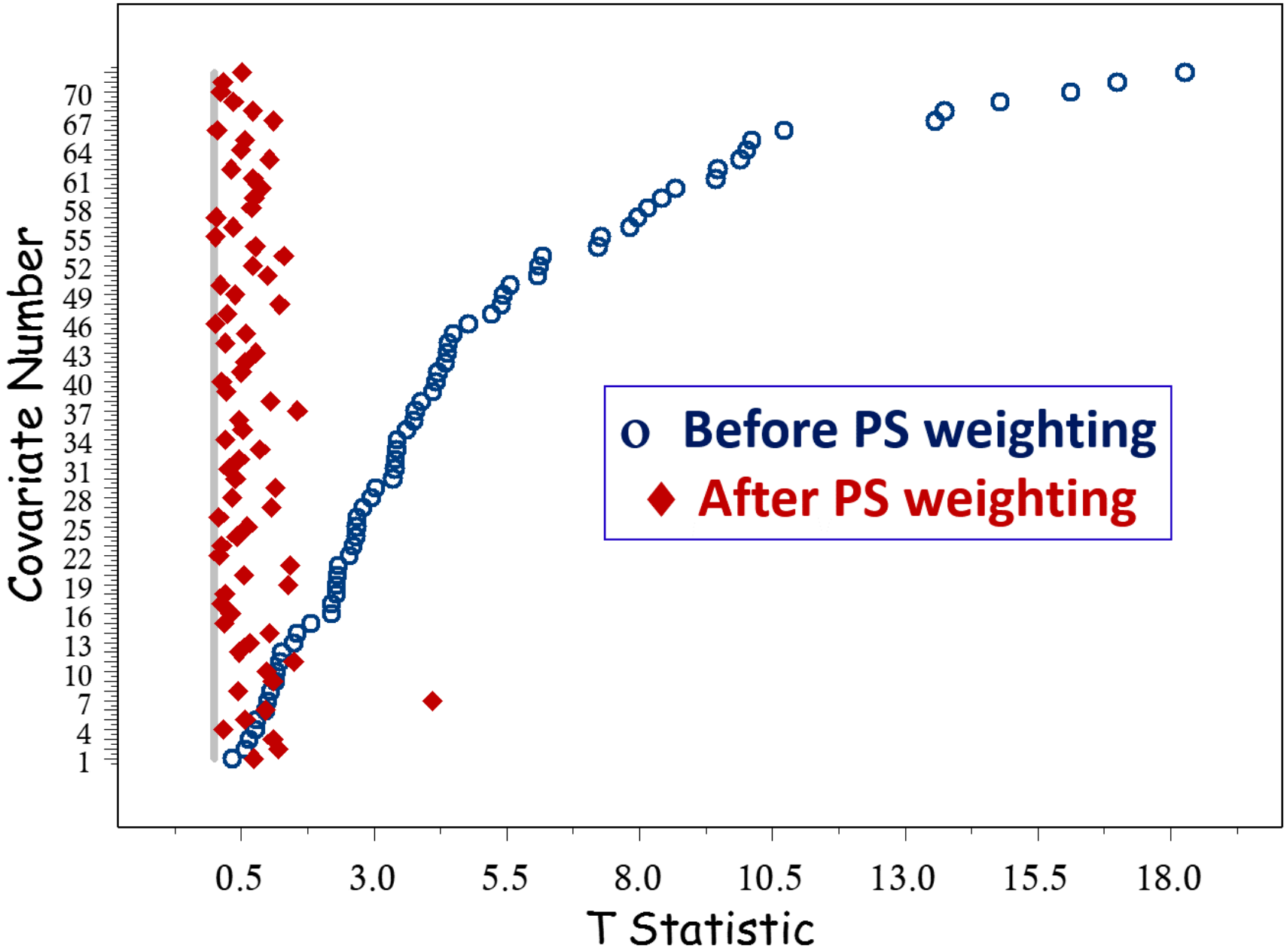
Reweight each treated patient by 1/PS, and each control patient by 1/(1-PS).

- PS model estimated by Hirano and Imbens[8] using 57 of 72 available covariates
  - Selected only those with $|t| > 2.0$
  - Serum potassium, for instance, prior to weighting showed a mean of 4.04 in the RHC group and 4.07 in the "No RHC" group, for a t = -0.99, so it was not included in the propensity model.

Results of this Weighting Approach on the next slide...

[8] Hirano and Imbens 2001, Connors 1996, Hirano, Imbens & Ridder 2003

Absolute T Statistics for RHC vs. No RHC Group Means



Effectiveness of RHC Propensity Score Weighting

- The weighting is based on a propensity model including 57 of the 72 covariates.
- Serum potassium not included in this PS.
- Most means are much closer, although six variables become less balanced (larger absolute t statistic) after weighting. None of these six were in the 57-variable PS model.
- Weighting by the propensity score appears to balance control and treatment groups well.

A "Double Robust" Estimator

1. Fit propensity score model
2. Weight the individual subjects (ATT, commonly) by the propensity score.
3. Directly adjust (via regression) for the propensity score in estimating the treatment effect.

- Forces you to think hard about selection.
- You don't care about parsimony in the PS, so you can maximize predictive value.
- Can fit a very complex PS model, and a smaller outcome model.
- Some hope that if PS model or weighting is helpful, the combination will be helpful.

Rosenbaum, Chapter 4

Adjustments for Measured Covariates

For Discussion

- What was the most **important** thing you learned from reading Chapters 1-3?
- What was the **muddiest**, least clear thing that arose in your reading?
- What questions are at the front of your mind now?

Section 7

Coming Next Time

Setting Up Class 04

The lecture will be a walk-through of the `toy` example, which is a simple simulated observational study of a treatment on three outcomes (one quantitative, one binary, and one time-to-event) which we will use to demonstrate the completion of 13 tasks using R, which include:

- Fitting a propensity score model
- Assessing pre-adjustment balance of covariates
- Estimating the effects of our treatment on our outcomes ...
  - Using matching on the propensity score
  - Using subclassification on the propensity score
  - Using direct adjustment for the propensity score
  - Using weighting on the propensity score

Note we have three other (more realistic) examples we'll share in time: `lindner`, `dm2200` and `rhc`.

Labs

- How did Lab 1 go?
- Lab 1 Sketch should be posted by Noon today.
- Lab 2 due 2024-02-15 at 9 AM to Canvas.

Progress on Semester Activities

- Searching for a suitable OSIA paper, and developing a claim by 2-6
- Building a proposal (first draft due 2-20) for the course project