500 Class 3 Slides

github.com/THOMASELOVE/500-2018

2018-02-01

Agenda for Class 3

- Tools for Assessing Causal Effects
  - Using Matched Sets to Adjust for Overt Bias
- Defining and Motivating the Propensity Score
- Rosenbaum, Chapters 5 and 6
  5. Between Observational and Experimental
  6. Natural Experiments
- Homework 2
- Using the Propensity Score

Returning to the Aspirin Example

Aspirin Use and Mortality (Gum 2001)

6174 consecutive adults at CCF undergoing stress echocardiography for evaluation of known or suspected coronary disease[1].

- 2310 (37%) were taking aspirin (treatment).
- Main Outcome: all-cause mortality
  - Median follow-up: 3.1 years
- Univariate Analysis: 4.5% of aspirin patients died, and 4.5% of non-aspirin patients died.
  - Unadjusted Hazard Ratio: 1.08 (0.85, 1.39)

[1] Gum PA et al. 2001

Matching on the Covariates, X

- We can create a **matched sample**, where we match treated subjects to control subjects, on the basis of their covariates.
  - Simplest is exact matching - but this can pose problems unless we have few covariates to deal with, with very limited possible values.
  - Often exact stratification or matching is impossible, but when it is, things go smoothly.

What's the difference between Aspirin Users and the other patients?

Variable	Aspirin Users	No Aspirin
Patients	2,310	3,864
Age, Mean (SD)	62 (11)	56 (12)
Male, %	77.0	56.1

Might it be reasonable to match up patients who are the same gender and similar in age? Or to stratify into groups by age and gender?

What's the difference between Aspirin Users and the other patients?

Variable	Aspirin Users	No Aspirin
Patients	2,310	3,864
Age, Mean (SD)	62 (11)	56 (12)
Male, %	77.0	56.1
Prior CAD, %	69.7	20.1
Beta Blocker, %	35.1	14.2

But now what do we do?

- How can we match on Age **and** Gender **and** history of CAD **and** beta-blocker prescription?
- Or (if that's not hard enough) how about the complete set of 31 covariates?

Using Matched Sets or Strata to Adjust for Overt Selection Bias

- Observe a set of p covariates, collected in $\mathbf{X}$
- Even if each covariate is binary, there are 2^p possible values of $\mathbf{X}$
  - Many subjects are likely to have unique values of $\mathbf{X}$.
- Realistic Goal: compare treated and control groups with similar distributions of $\mathbf{X}$, even if matched individuals have differing values of $\mathbf{X}$

Key tool for doing this well: propensity score

What Do We Want to Know about a Clinical or Health System Intervention[2]?

- Response: Can we estimate the impact of the intervention? Can we estimate costs and benefits?
- Predictors: Can we "mine" for attributes that help predict response to the intervention?
- Evaluation: Can we fairly estimate the average health impact of our intervention?
- Target Evaluation: Can we identify likely responders? Subgroup analyses?

[2] from a marketing list at www.anabus.com

The Data You Wish You Had

Subject	Health if exposed	if unexposed
A	12	8
B	7	4
C	7	3
D	12	9

ALL potential outcomes available!

The Data You Wish You Had

Subject	Health if exposed	if unexposed	Exposure Effect
A	12	8	4
B	7	4	3
C	7	3	4
D	12	9	3

Wouldn't this be great!

Grim Reality

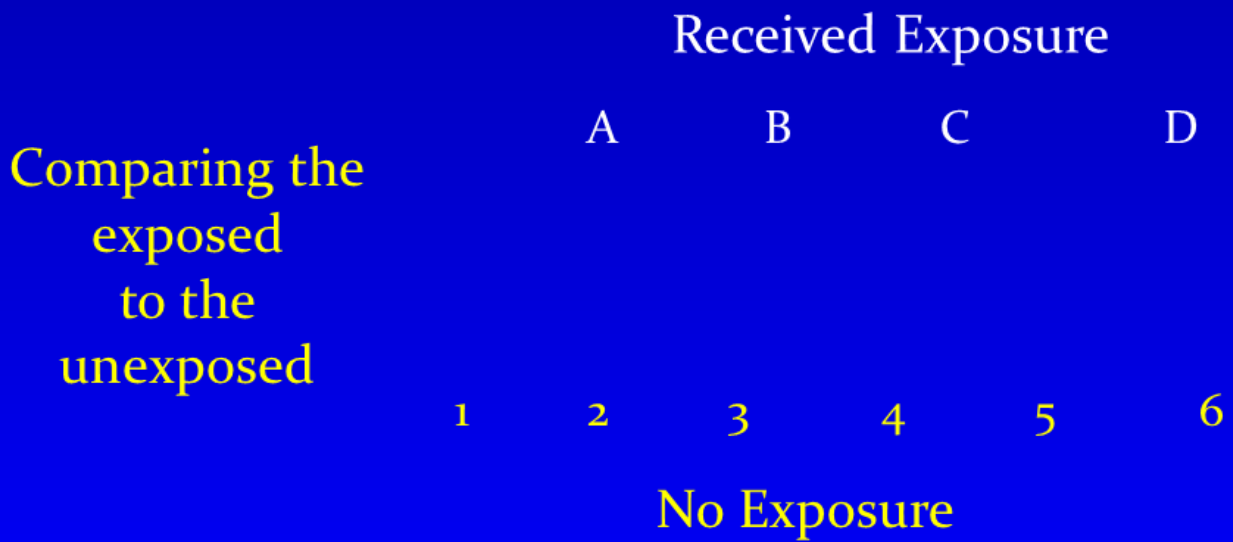
Subject	Health if exposed	if unexposed	Exposure Effect
A	12	?	?
B	7	?	?
C	?	3	?
D	?	9	?

Causal inference is a **missing data** problem.

How should we fill in those question marks?

Matching and Causal Effects

Causal Analysis



Causal Analysis

Causal Analysis

Received Exposure

A B C D

Add baseline characteristics

1 2 3 4 5 6

No Exposure

Causal Analysis

Received Exposure

A B C D

Add baseline characteristics

1 2 3 4 5 6

No Exposure

Causal Analysis

Received Exposure


A


B


C


D

Add baseline characteristics

1 2 3 4 5 6

No Exposure

Causal Analysis

Received Exposure

A B C D

Add baseline characteristics

1 2 3 4 5 6

No Exposure

Causal Analysis

Received Exposure

A B C D

Which pairs should we compare?

1 2 3 4 5 6

No Exposure

Causal Analysis

Received Exposure

A B C D

Which pairs should we compare?

1 2 3 4 5 6

No Exposure

Causal Analysis

"Comparing Apples to Apples"

Received Exposure

A B C D

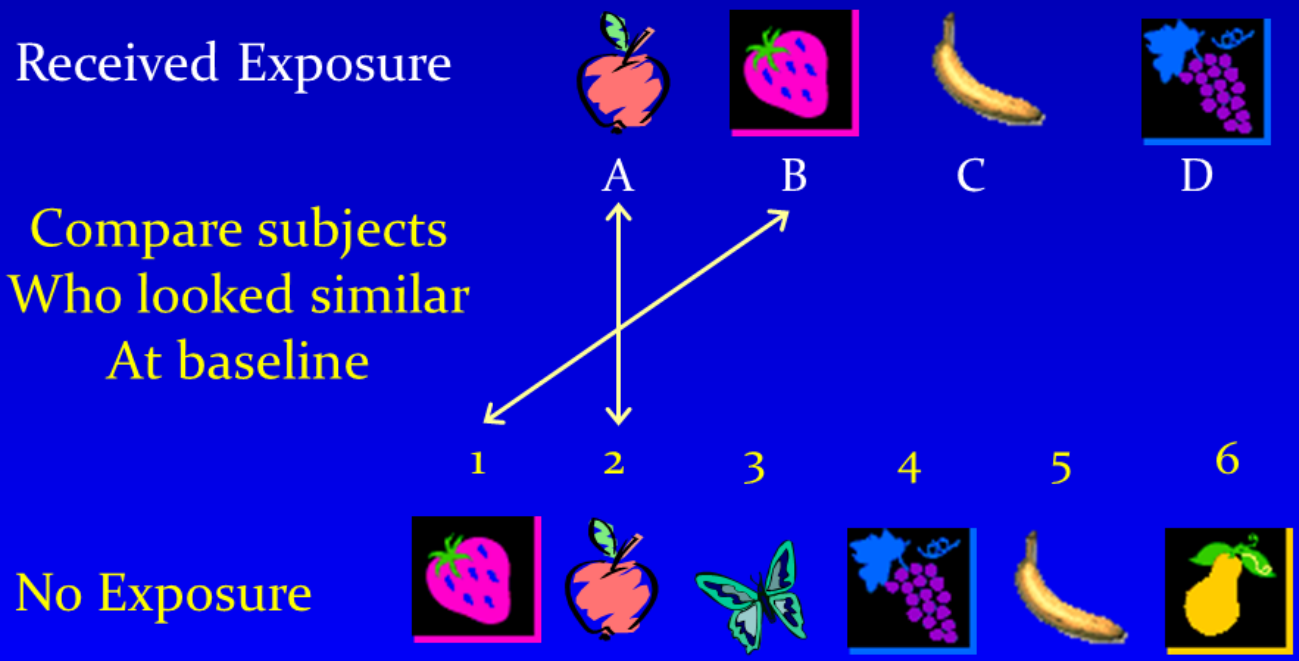
Which pairs should we compare?

1 2 3 4 5 6

No Exposure

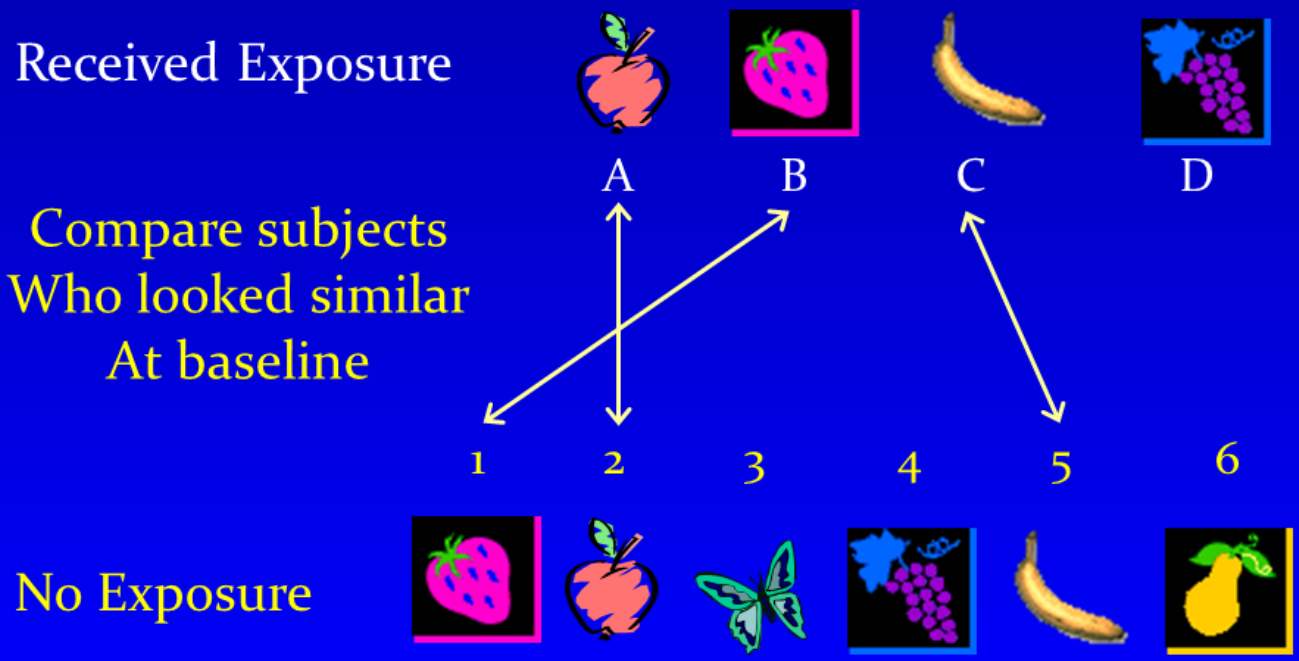
Causal Analysis

"Comparing Apples to Apples"



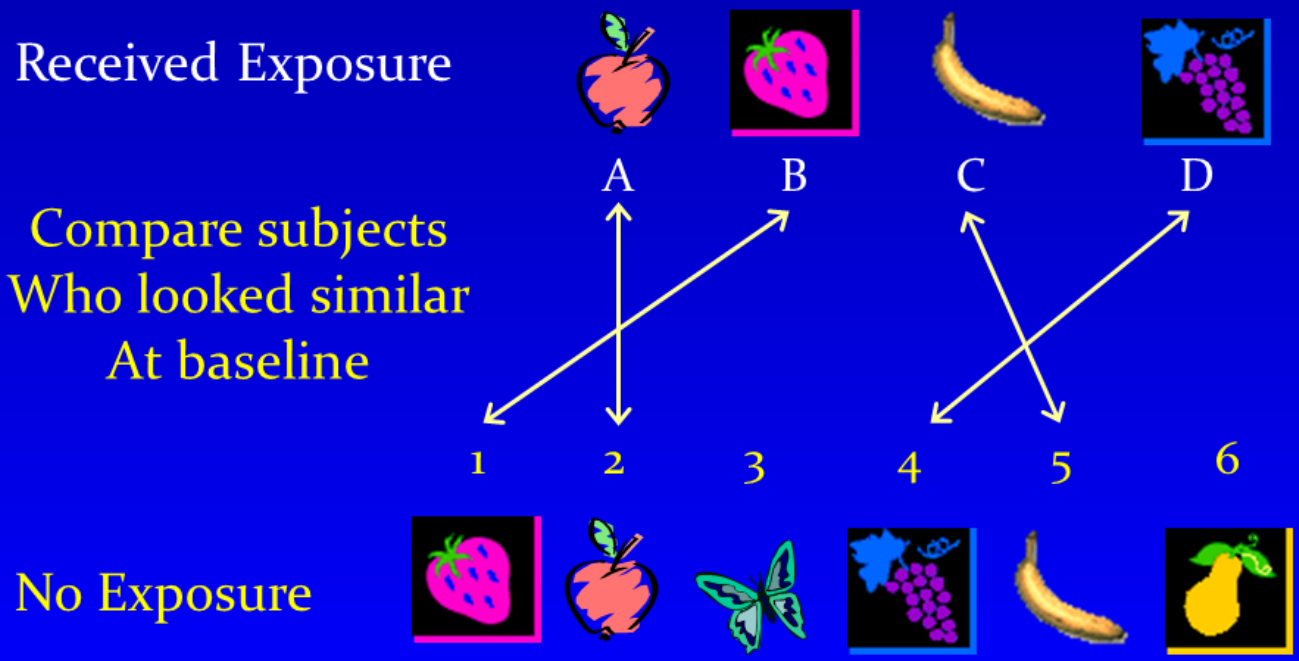
Causal Analysis

"Comparing Apples to Apples"



Causal Analysis

"Comparing Apples to Apples"



The Propensity Score

The Propensity Score

Definition: The conditional probability that a subject receives an exposure given the values of their vector of covariates.

- PS = Pr(exposed | covariates)

Reduces the baseline information to a single, composite summary of the covariates, between 0 and 1.

- Of course, we know whether a subject in fact either receives or doesn't receive the exposure.
- But we will estimate this probability for each subject as a convenient way of expressing the impact of covariate information on the exposure assignment decision, as a scalar value between 0 and 1.

Estimating the Propensity Score (most common approach)

Estimate a Logistic Regression Model:

- Y = Exposure Group
  - 1 = exposed, 0 = unexposed
- Predictors are the observed covariates

Use anything related to exposure decisions that can be collected prior to exposure assignment.

Propensity Scores = Predicted Pr(exposure) for each subject, i.e. the **fitted values**

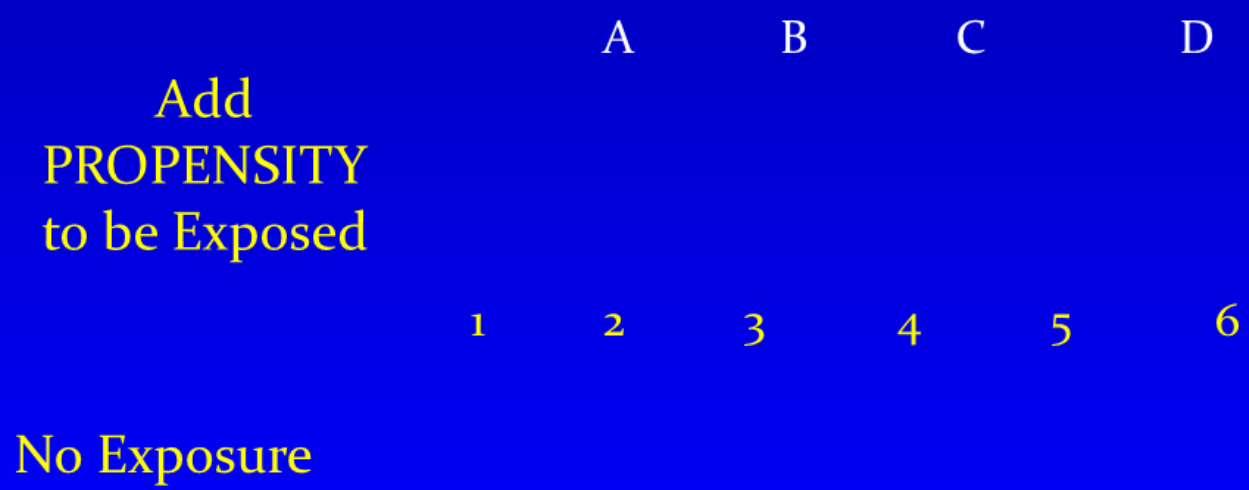
Why Estimate the Probability that a subject was "exposed"?

- Using Pr(subject would have been exposed), we create a quasi-randomized experiment.
- If we have two subjects, one treated and one control, with the same propensity score, we can imagine that these two subjects were randomly assigned to each group - just as if we were doing an experiment!
- Except that here we can't assume that we control for anything that we didn't measure.

Causal Analysis

Causal Analysis

Received Exposure .62



Causal Analysis

Received Exposure	.62	.74	.59	.81
	A	B	C	D

Add
PROPENSITY
to be Exposed

1 2 3 4 5 6

No Exposure

Causal Analysis

Received Exposure

.62	.74	.59	.81
A	B	C	D

Add PROPENSITY to be Exposed

1	2	3	4	5	6
.74	.62	.36	.80	.58	.23

No Exposure

Causal Analysis

Received Exposure

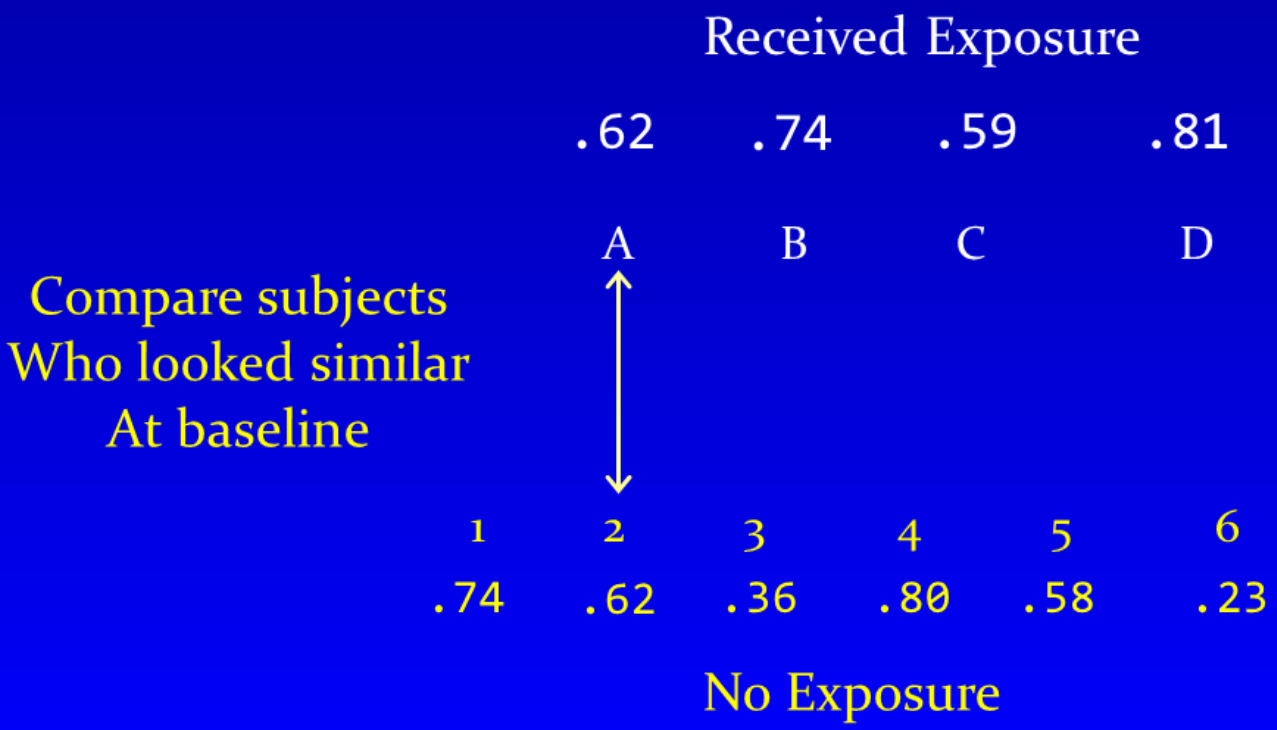
.62	.74	.59	.81
A	B	C	D

Which pairs should we compare?

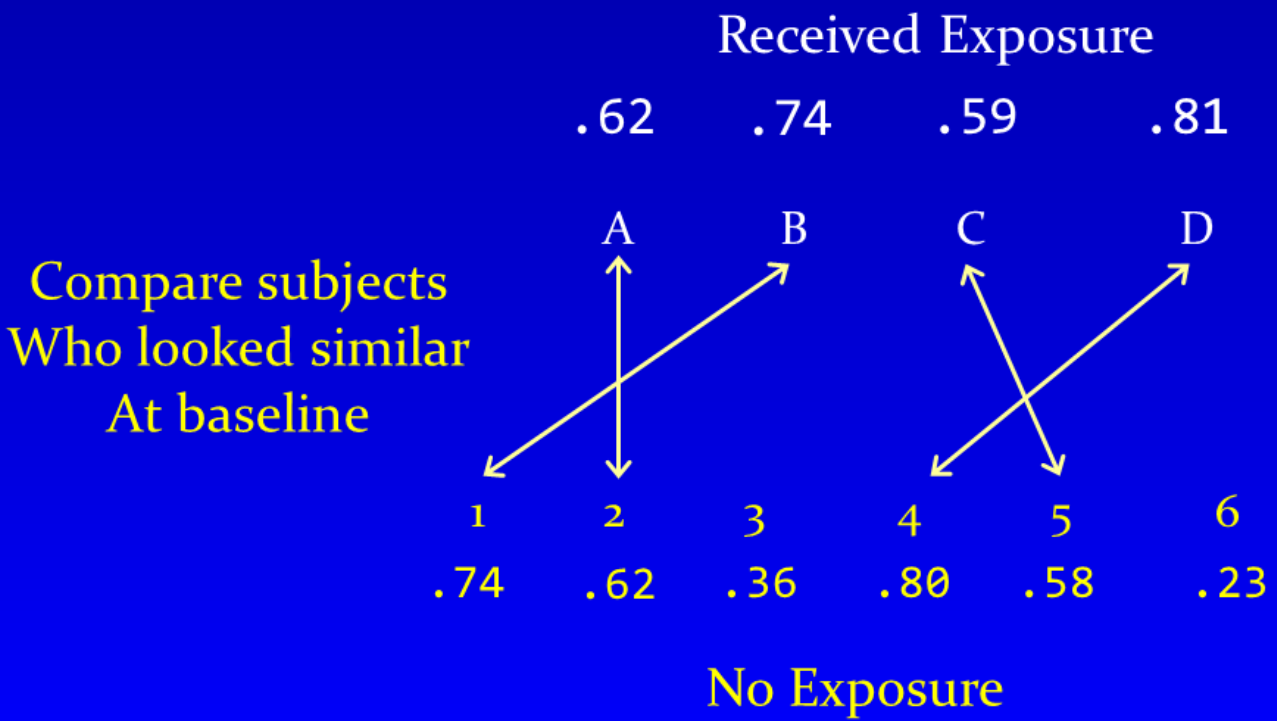
1	2	3	4	5	6
.74	.62	.36	.80	.58	.23

No Exposure

Causal Analysis



Causal Analysis



Grim Reality

Subject	Health if exposed	Health if unexposed	Exposure Effect
A	12	?	?
B	7	?	?
C	?	3	?
D	?	9	?

Improving Grim Reality

Subject	Propensity for Exposure	Health if exposed	if unexposed
A	0.80	12	?
B	0.50	7	?
C	0.51	?	3
D	0.79	?	9

- Can we use the propensity score to guide our matching approach?
- Can we plug in resulting estimates for our question marks?

Propensity Score Matching yields a new Database

Subject	PS	Health if exposed	if unexposed	Exposure Effect
A	0.80	12	[9]	[3]
B	0.50	7	[3]	[4]
C	0.51	[7]	3	[4]
D	0.79	[12]	9	[3]

Now, we can estimate the **impact of the exposure** on each matched patient.

How Do We Use the Propensity Score?

1. Start with a sample where the exposed subjects don't look very much like the unexposed subjects.
2. Adjust the sample (in some manner) to make the distributions of exposed and unexposed subjects look more similar prior to exposure.
3. This will let us attribute the differences we see in outcomes between these adjusted samples more easily to the exposure's causal effect, and not so much to the original differences between the groups.

- To do this, we estimate the propensity score: the probability of receiving the exposure for each subject given their covariate values.
- Then, we use the propensity score in one of the ways listed on the next slide to fuel our estimates of causal effects.

Methods for Using Propensity Scores

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

All of these are found in the `toy` example on our web site.

Rosenbaum Chapters 5-6

Rosenbaum, Chapters 5-6

5. Between Observational and Experimental
6. Natural Experiments

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

Homework 2 discussion

Homework 2

Due at noon on 2018-02-01.

- Any problems?

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 $\mathbf{X}$) but not (directly) for the potential biasing effects of omitted/hidden covariates
- often, but not inevitably, fit with a "kitchen sink" logistic regression[3]

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

[3] 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[4]?
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?

[4] see Frank Harrell reference

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.

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.

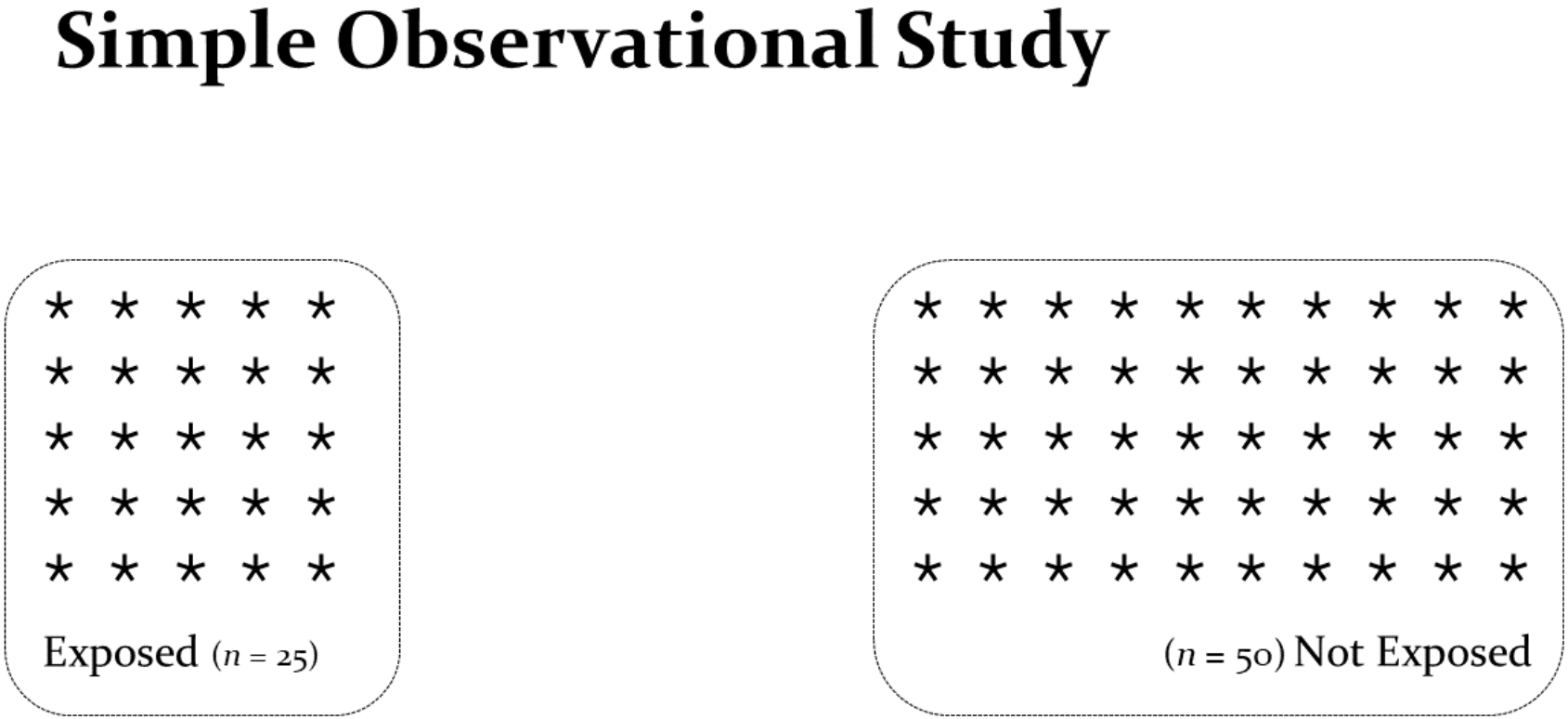
Should we be checking propensity model goodness of fit?

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

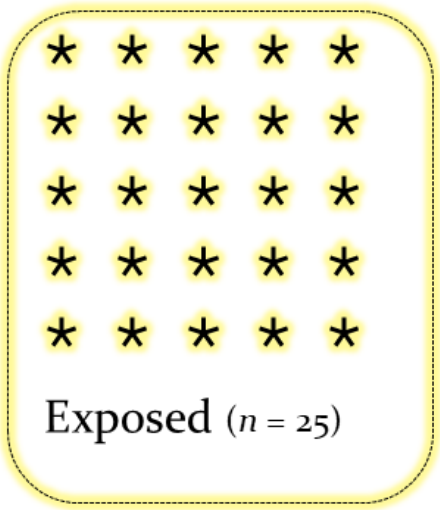
- Simulated data including an important binary confounder, and they compared inclusion to exclusion
- Hosmer-Lemeshow GOF test and C statistic were of no value in detecting residual confounding in treatment effect estimates

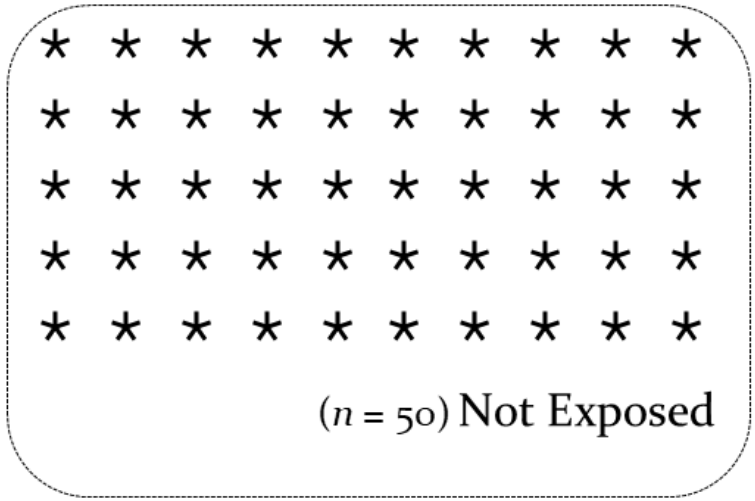
Using the Propensity Score: Some Schematics

A Simple Observational Study

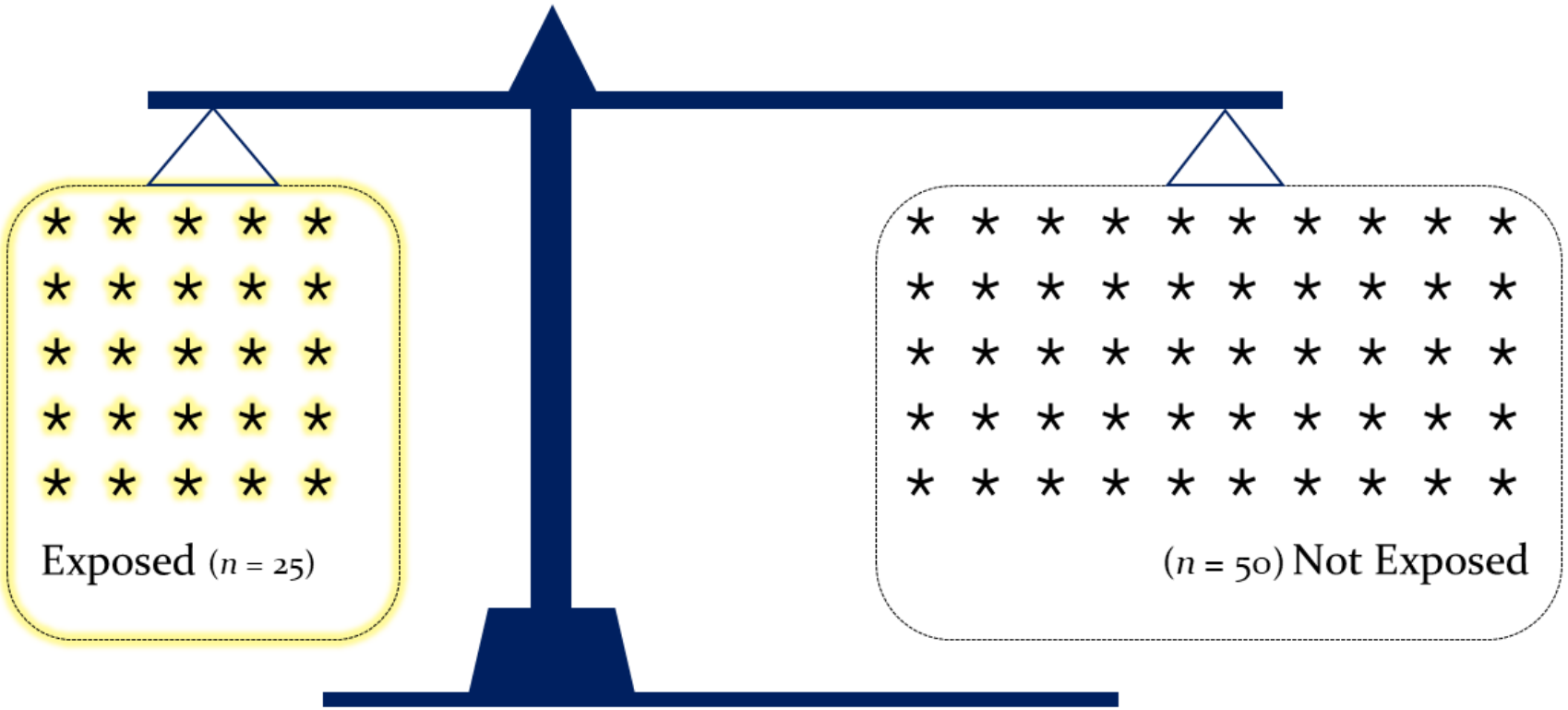


Apply the Exposure





To estimate causal effects, we need the baseline covariates to be in balance...



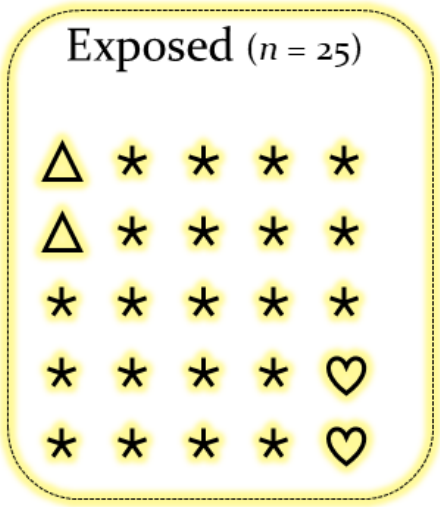
Model Without the Propensity Score

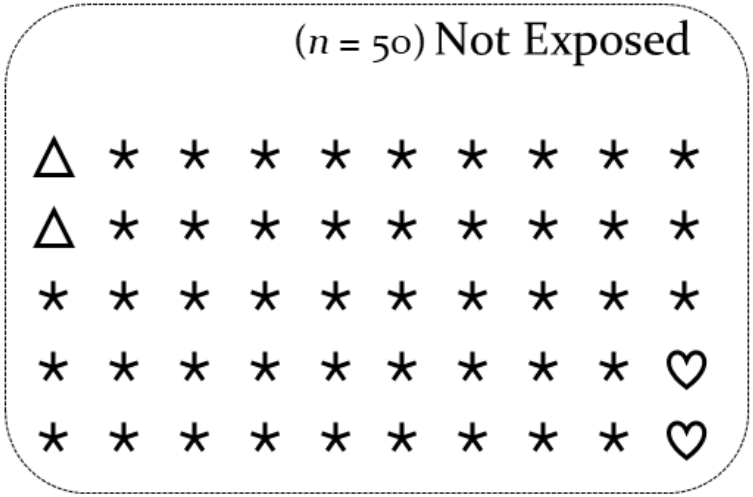
Outcome = $ß_0$ + $ß_1$*Exposure,

for pool of 75 subjects

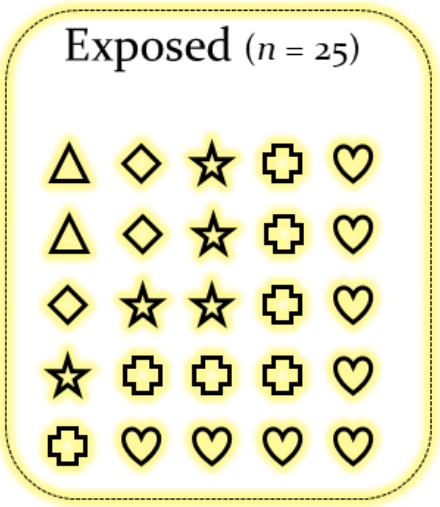
We interpret $ß_1$ as the exposure's effect.

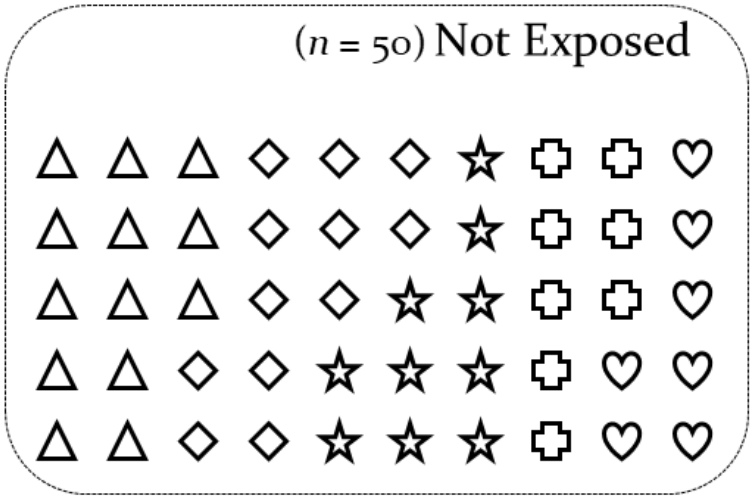
Subjects vary, within exposure groups...



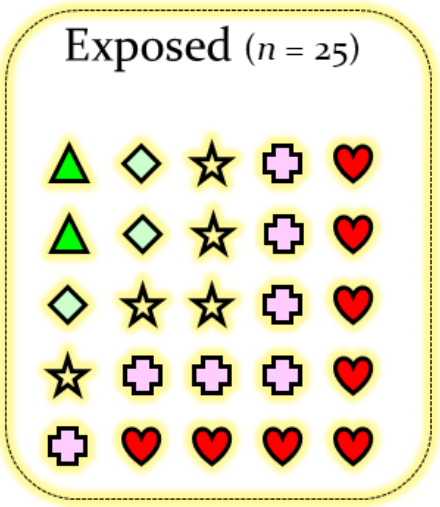


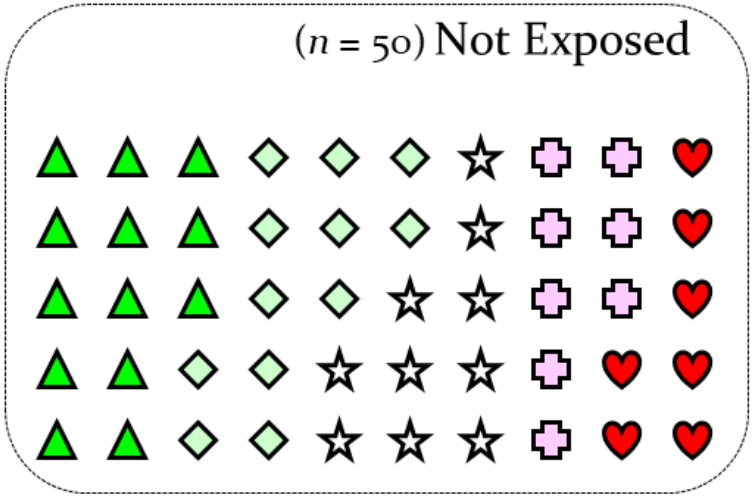
Actually, they vary quite a bit ...



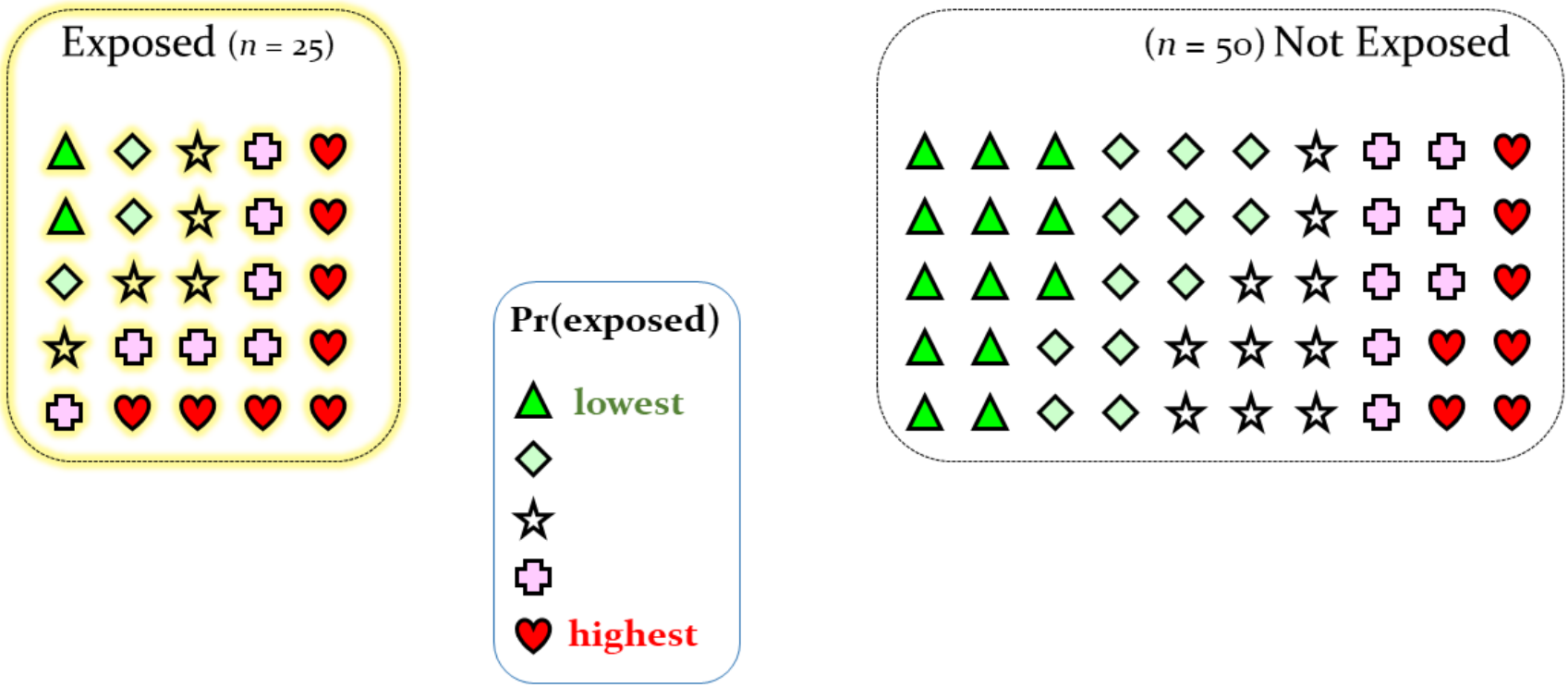


Actually, they vary even more than that

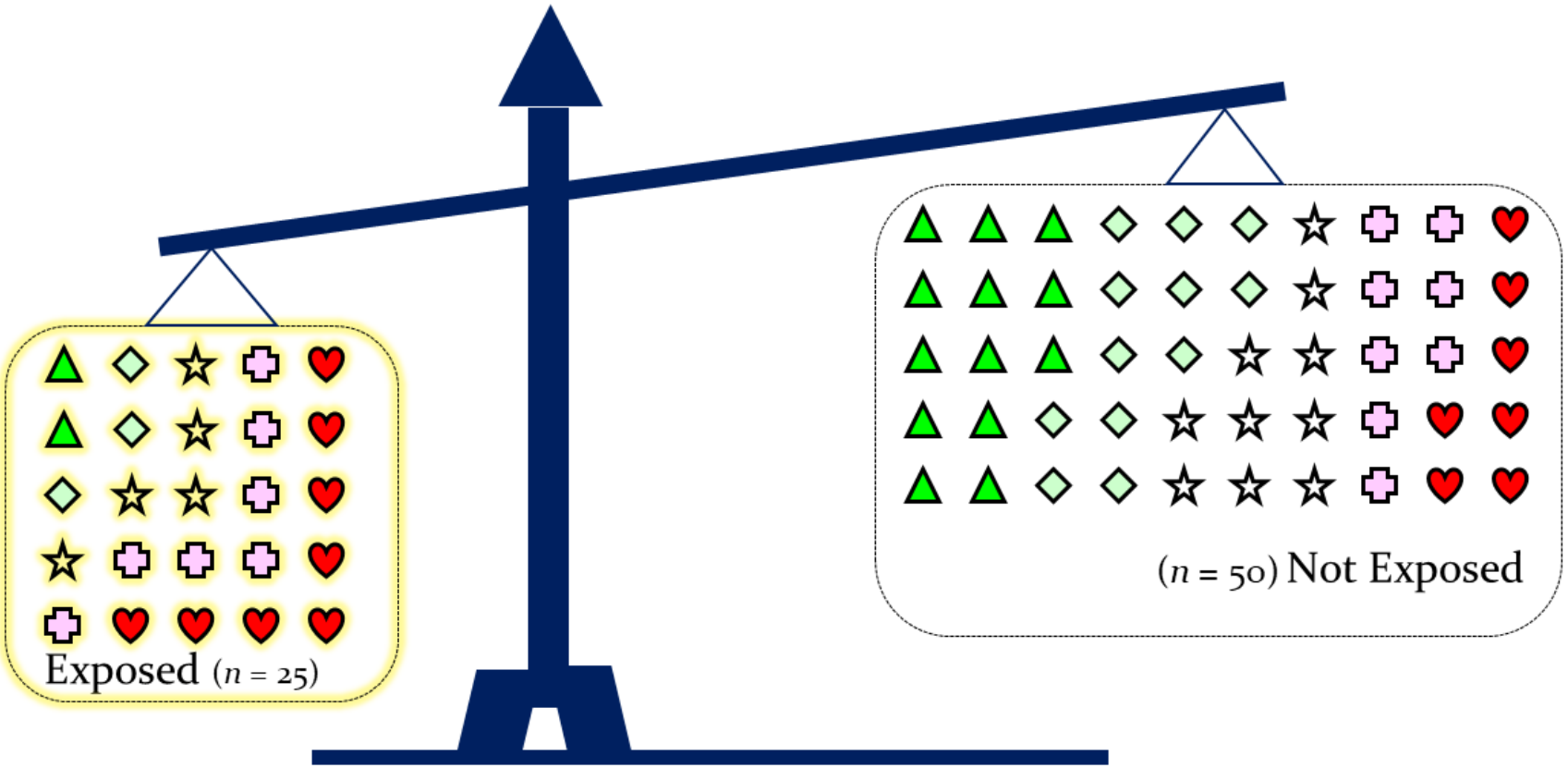




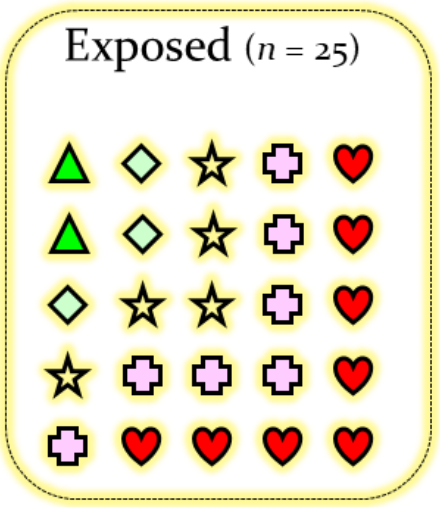
Characterize by propensity to receive the exposure...

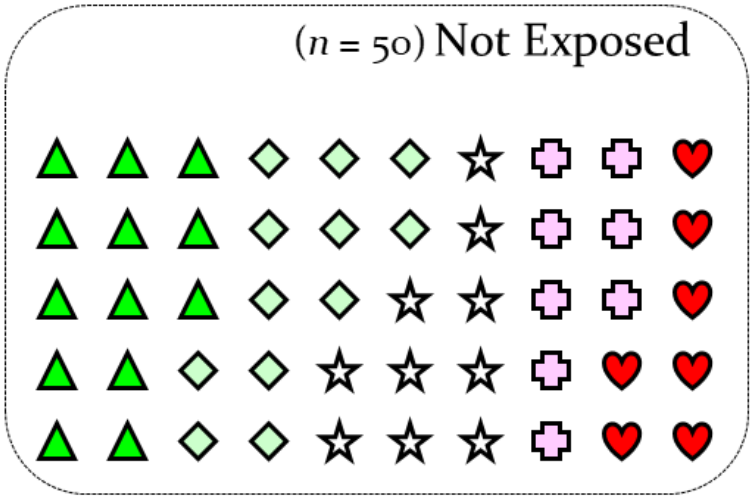


Are baseline characteristics in balance?



Comparing Exposure Groups Fairly?



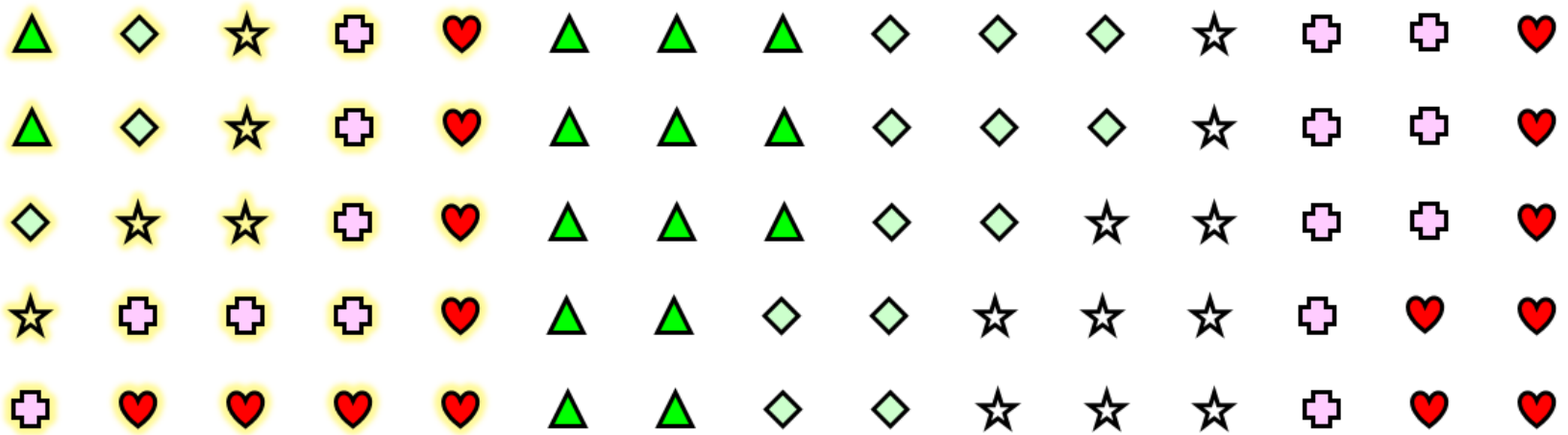


Model Without the Propensity Score

Outcome = $ß_0$ + $ß_1$*Exposure,

for pool of 75 subjects

We interpret $ß_1$ as the exposure's effect.

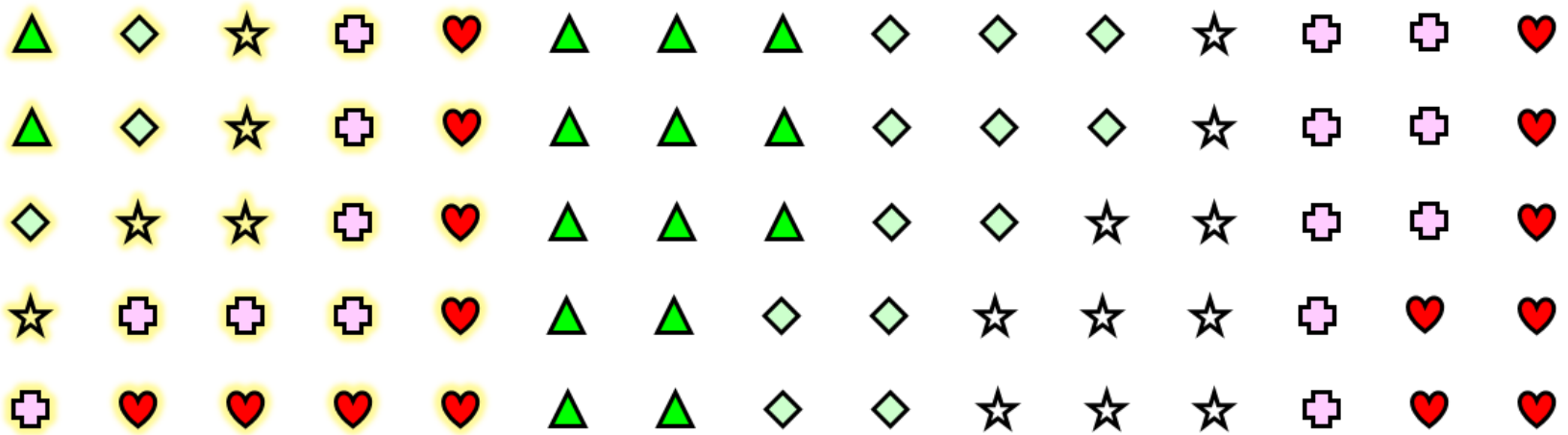


Could include covariates, as well...

$$\text{Outcome} = \beta_0 + \beta_1 * \text{Exposure} + \beta_j * \text{Covariate}_j,$$

for pool of 75 subjects

Still interpret β_1 as the exposure's effect, after covariate adjustment.



Propensity Score Subclassification / Stratification

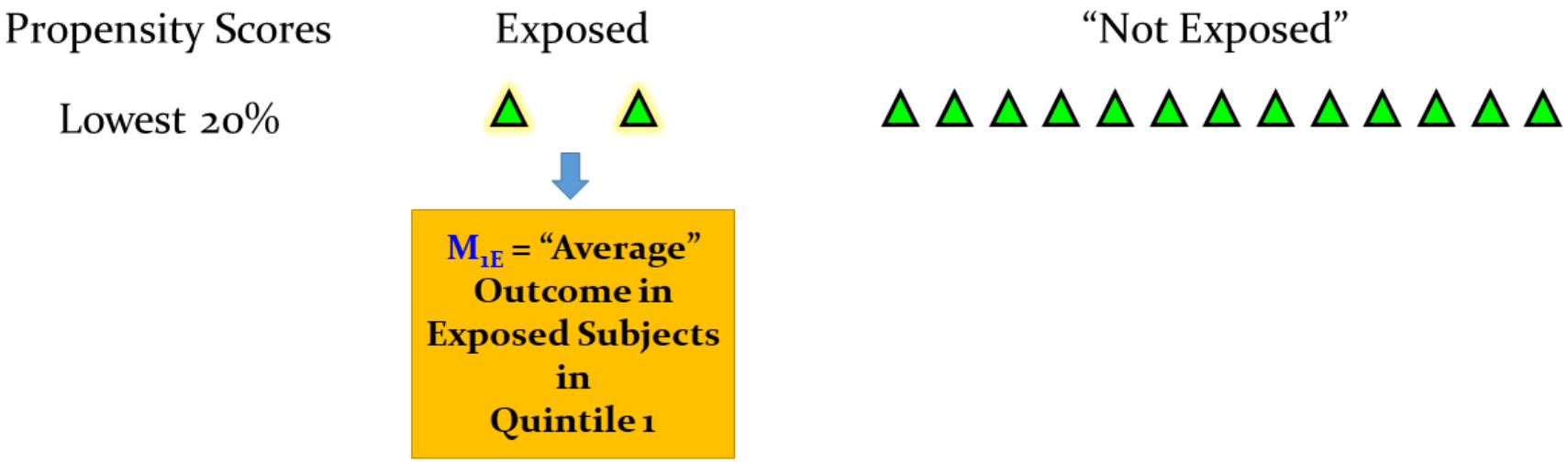
Propensity Score Subclassification

Propensity Score Subclassification

1. Split Subjects by Propensity Quintile

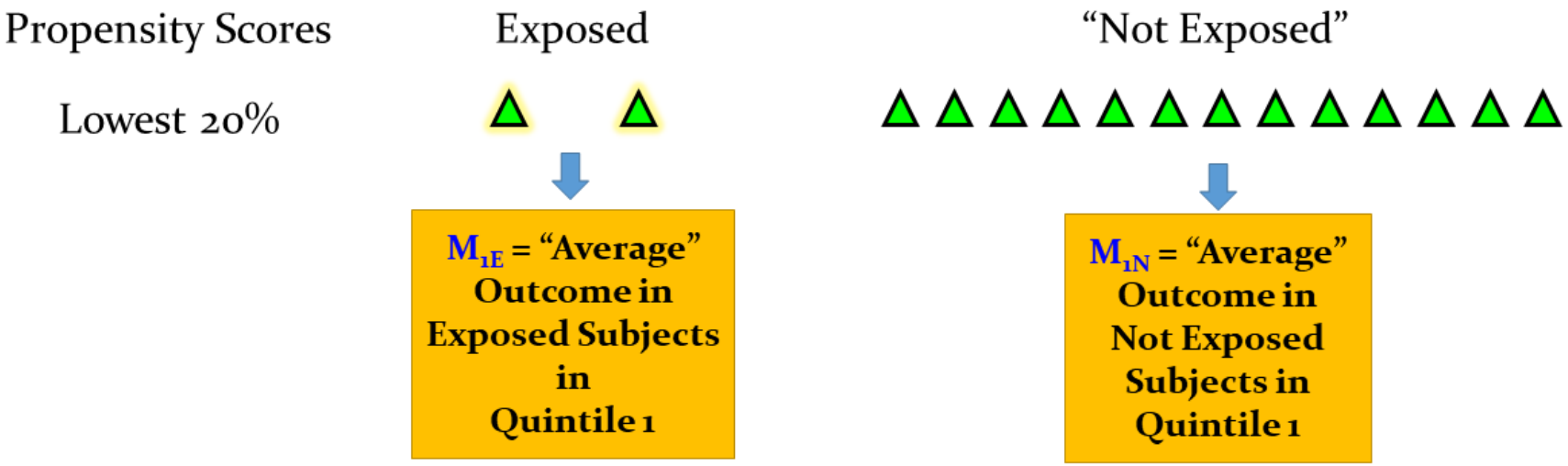
Propensity Score Subclassification

2. Estimate Effects Separately by Quintile



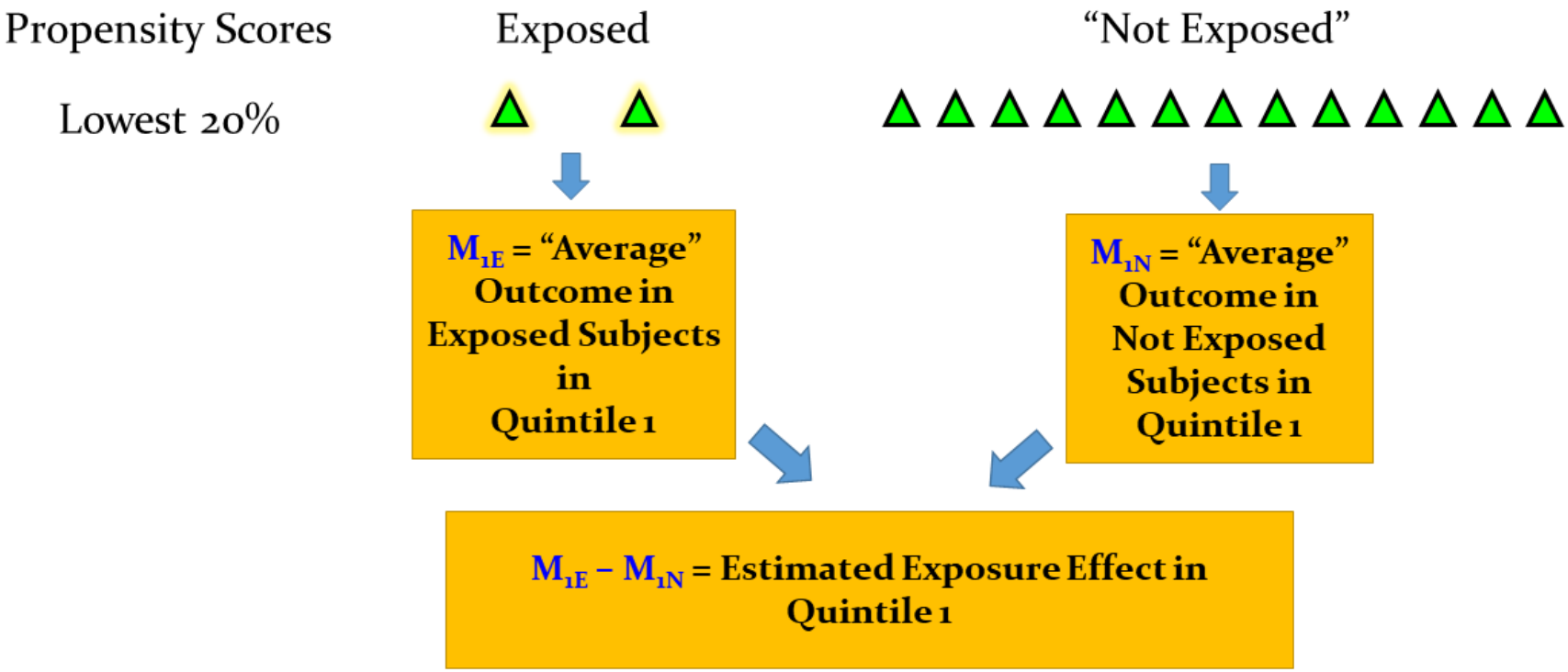
Propensity Score Subclassification

2. Estimate Effects Separately by Quintile



Propensity Score Subclassification

2. Estimate Effects Separately by Quintile



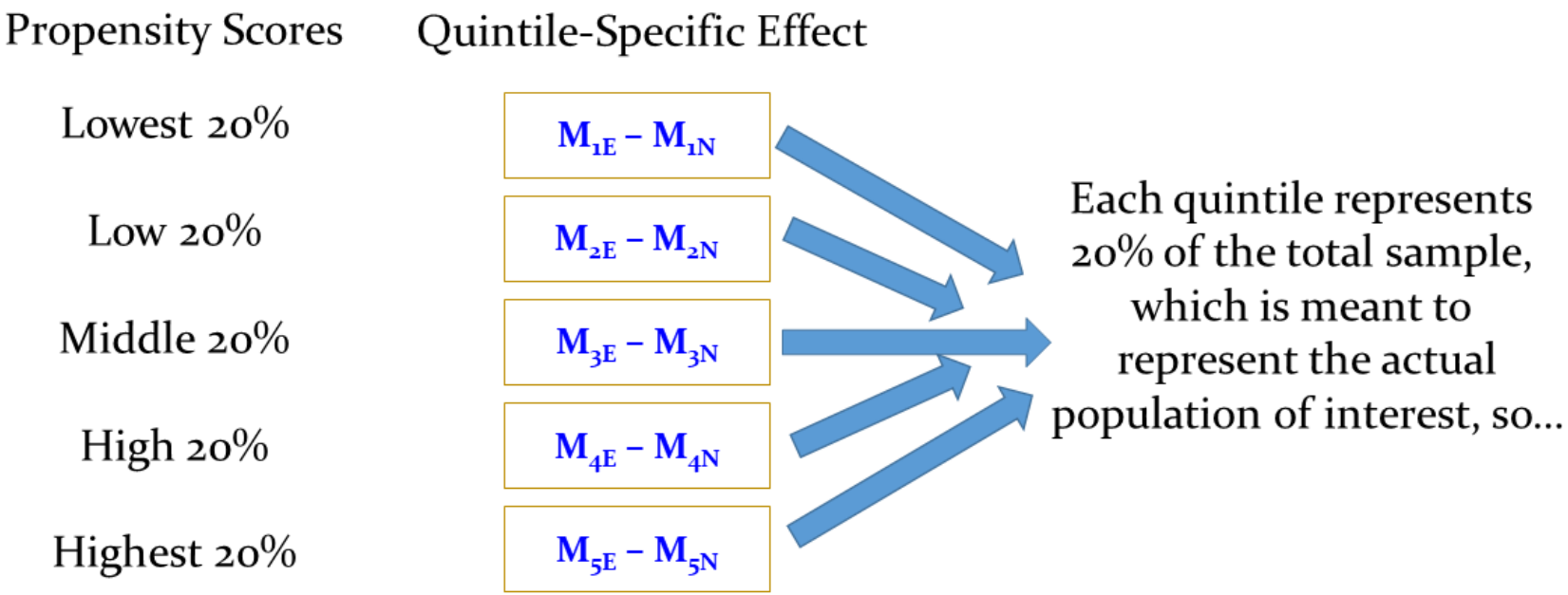
Propensity Score Subclassification

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

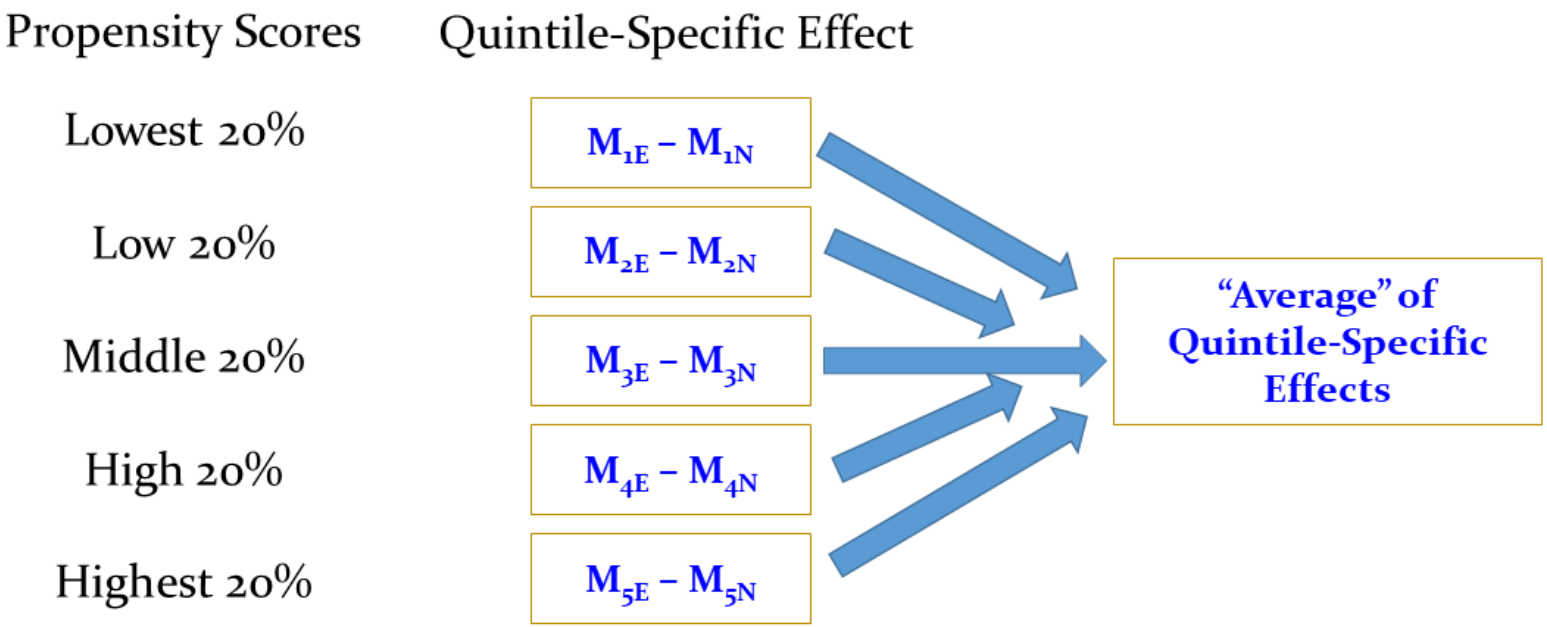
Propensity Score Subclassification

3. Combine Quintile-Specific Estimates



Propensity Score Subclassification

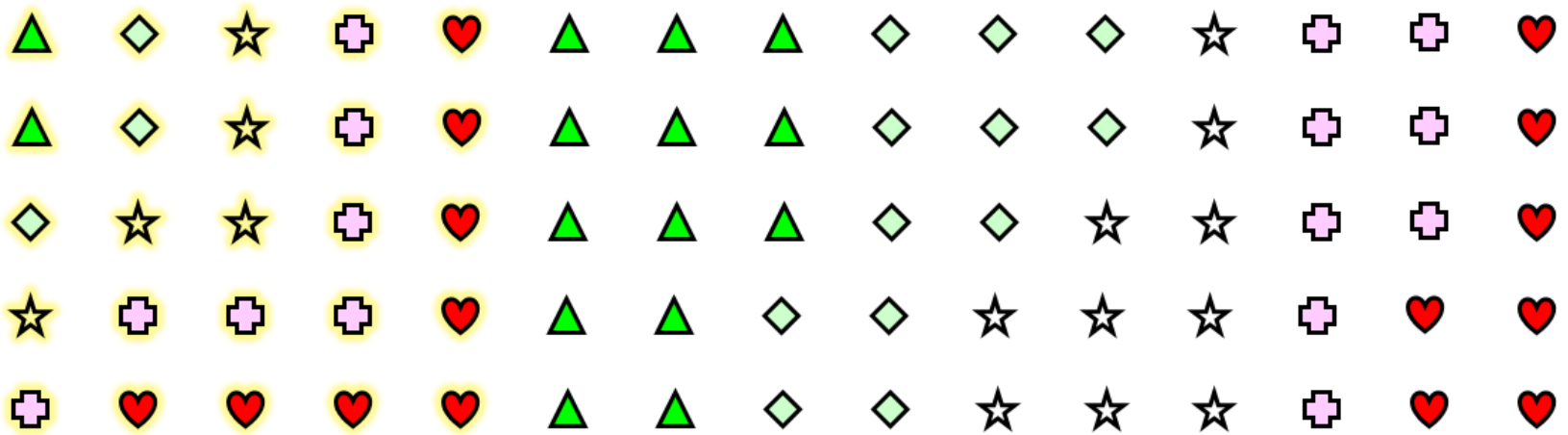
3. Combine Quintile-Specific Estimates



Direct Adjustment for the Propensity Score

Model Without the Propensity Score

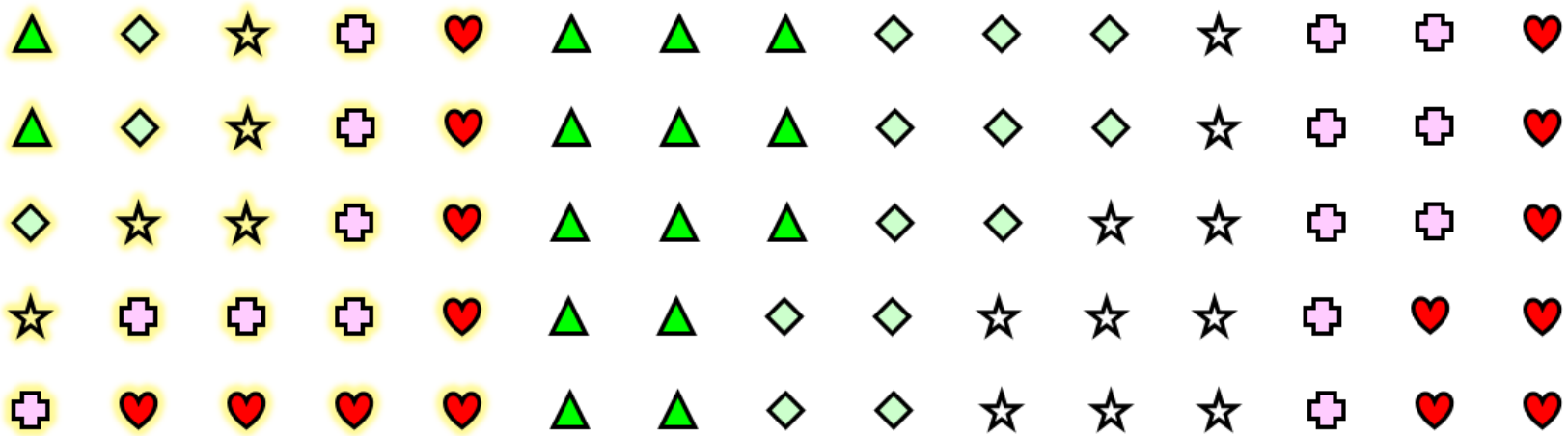
Outcome = $ß_0$ + $ß_1$*Exposure, for pool of 75 subjects



Direct Adjustment for Propensity Score

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

Again, across entire pool of 75 subjects



Propensity Score Weighting

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

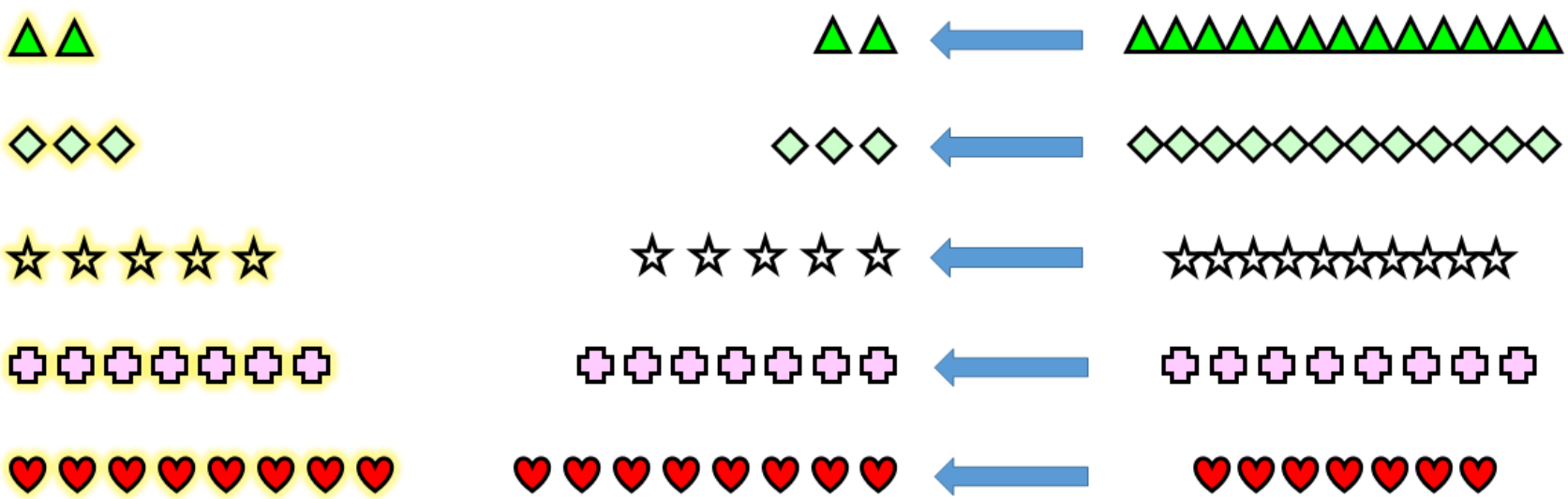
"Not Exposed"
unweighted

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)

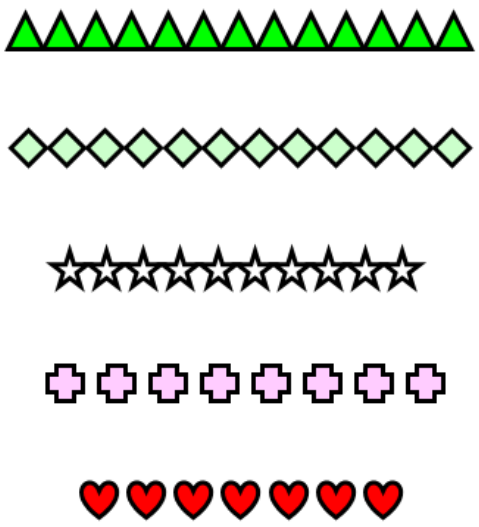
"Weighted Average"
Effect of Exposure on
Outcome

Propensity Score Matching

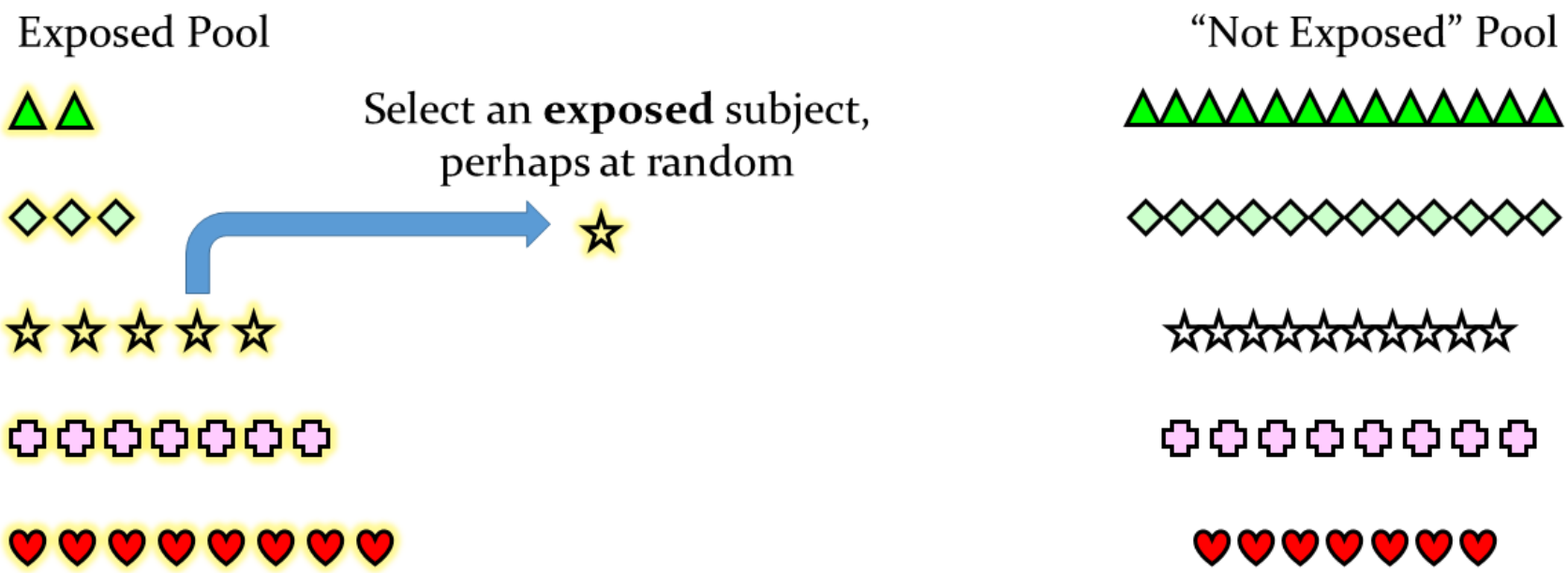
Propensity Score Matching (1:1)

Exposed Pool

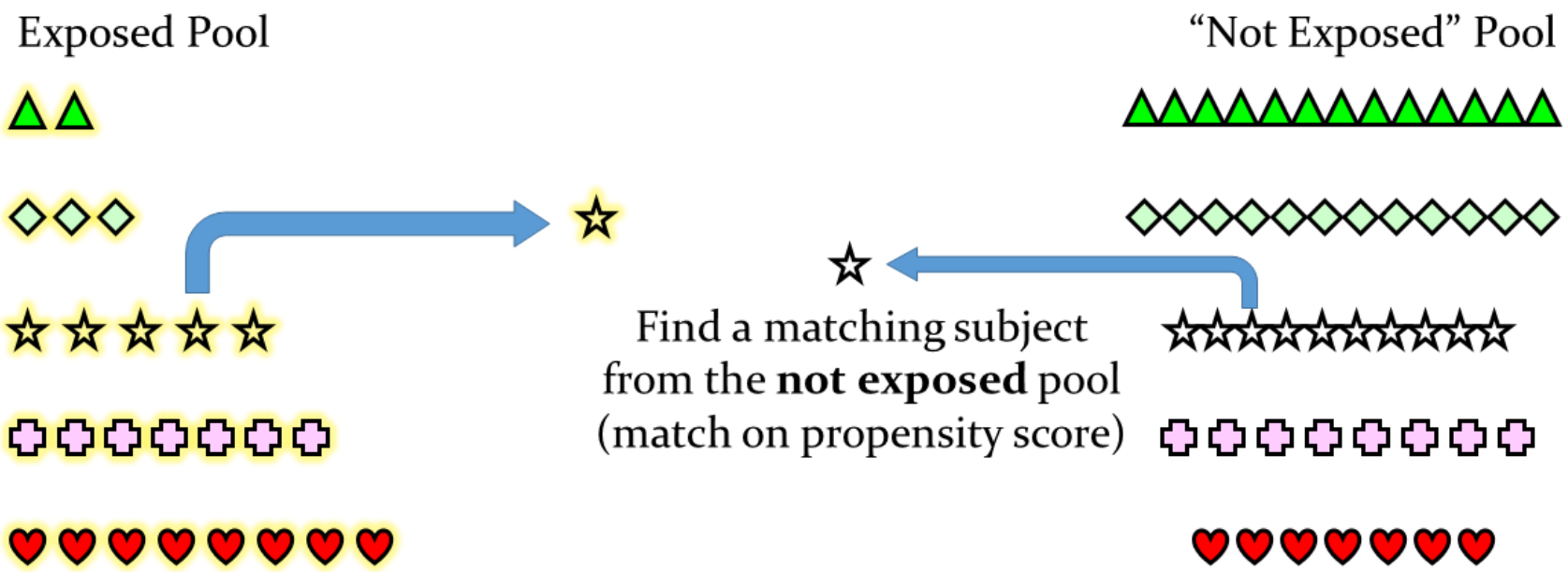
"Not Exposed" Pool



Propensity Score Matching (1:1)



Propensity Score Matching (1:1)



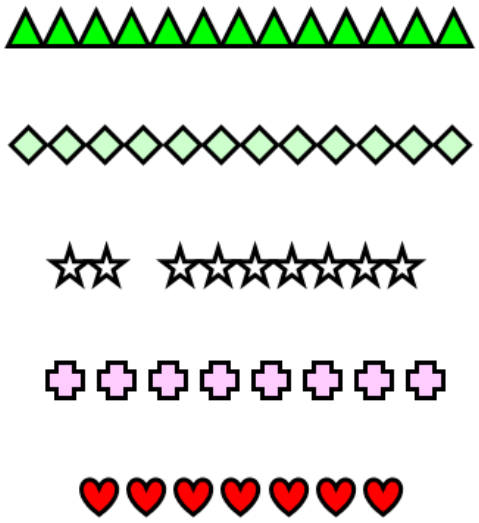
Propensity Score Matching (1:1)

Exposed Pool

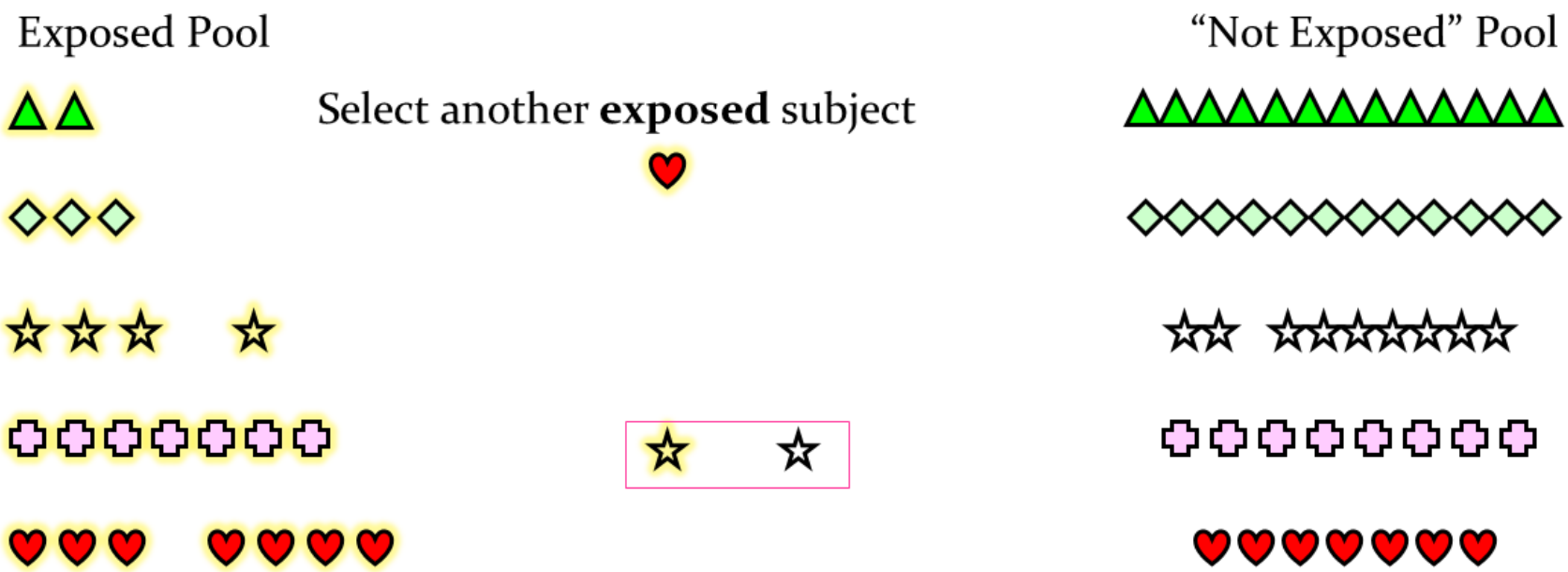
Form a matched pair

We're matching without replacement.

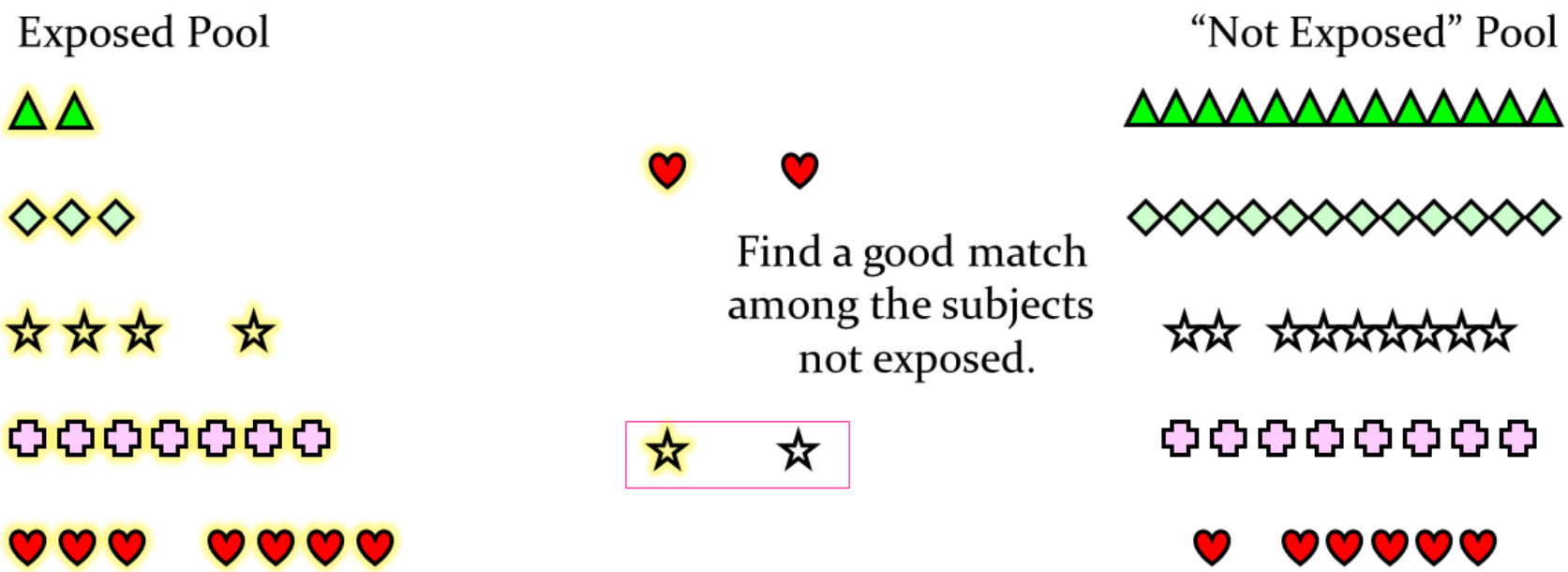
"Not Exposed" Pool



Propensity Score Matching (1:1)



Propensity Score Matching (1:1)

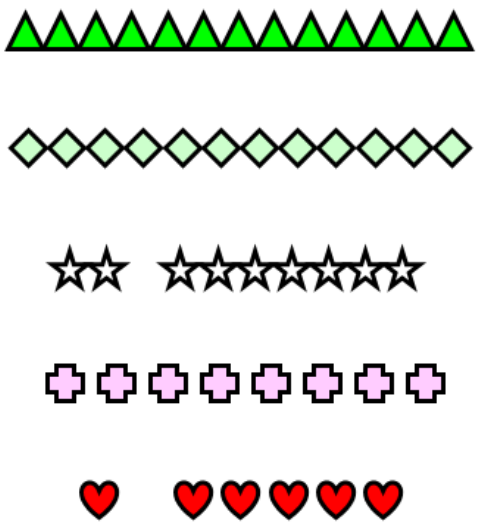


Propensity Score Matching (1:1)

Exposed Pool

A second matched pair!

"Not Exposed" Pool

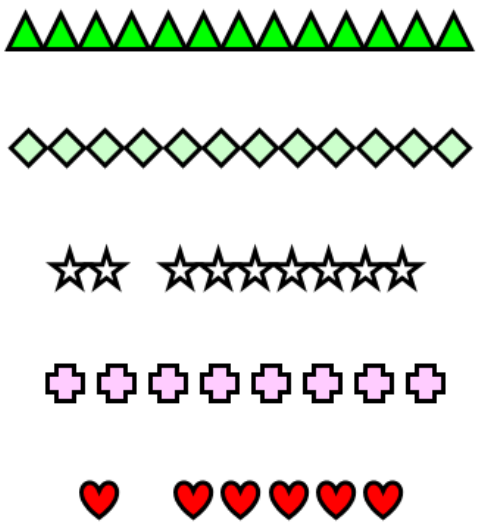


Propensity Score Matching (1:1)

Exposed Pool

Keep matching, until we can find no more acceptable matches

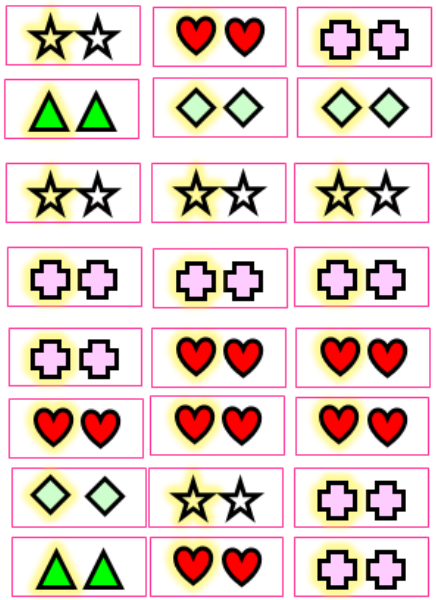
"Not Exposed" Pool



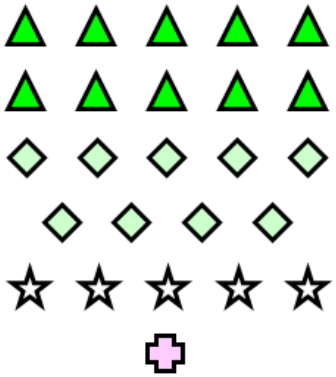
Propensity Score Matching (1:1)

Exposed Pool
(unmatched)

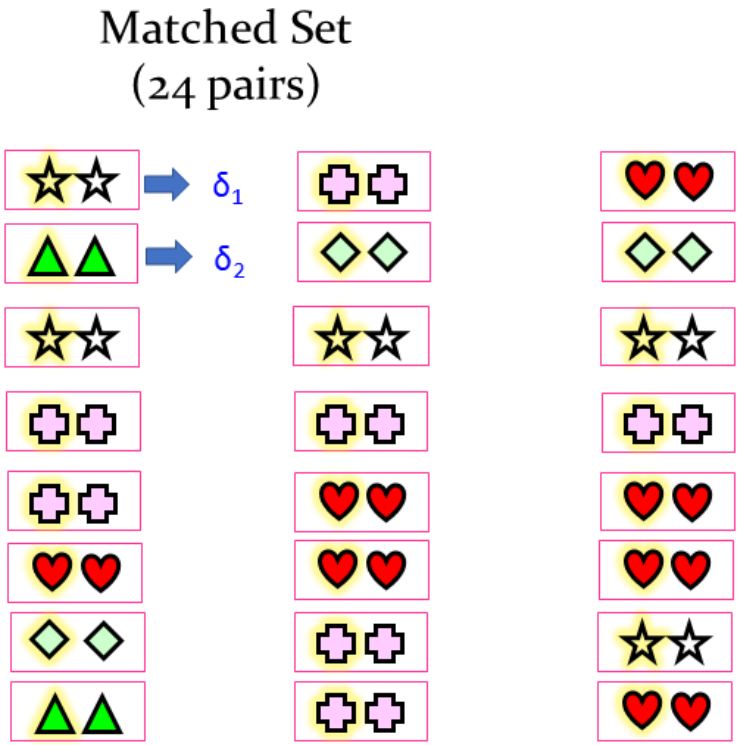
Matched Set
(24 pairs)



"Not Exposed" Pool
(unmatched)



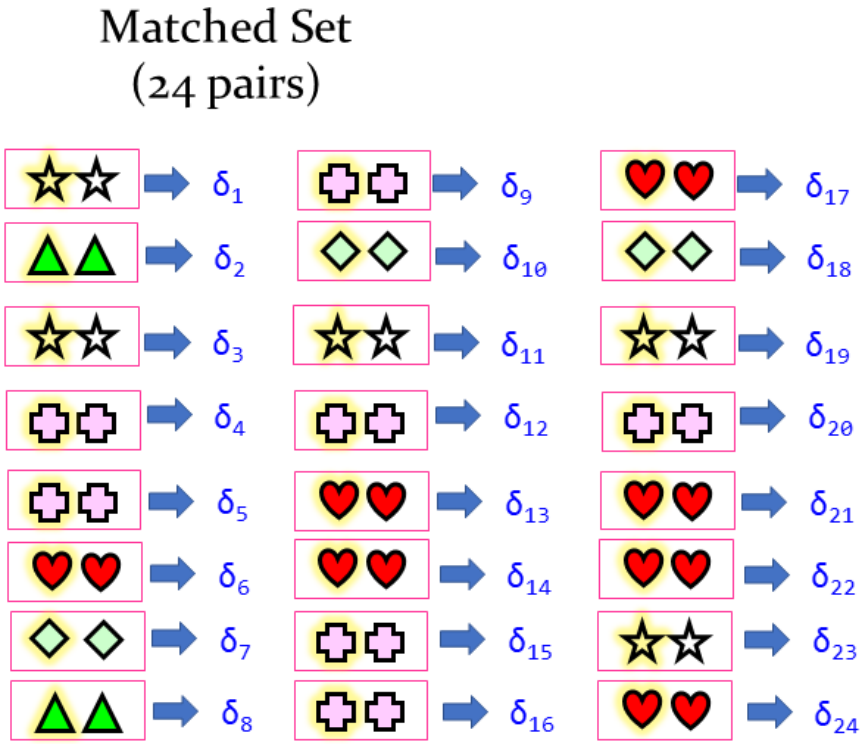
Propensity Score Matching (1:1)



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)



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 Propensity Scores Can and Cannot Do

- If we match (subclassify) treated subjects to controls with similar propensity scores, we can behave as if they had been randomly assigned to treatments.
- Or, if we use regression to adjust for propensity to get treatment, we can compare treated to controls without worrying about the impact of baseline differences we've measured on selection to exposure.
- But if our propensity model misses an important reason why subjects are selected to an exposure, we'll be in trouble, and **never know it**.

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[5].

- 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)

[5] Seminal paper: Rosenbaum and Rubin (1983)

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[6].

- 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).

[6] 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.

.80
.79
.78
.77
.76
.75
.74
.73
.72

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)

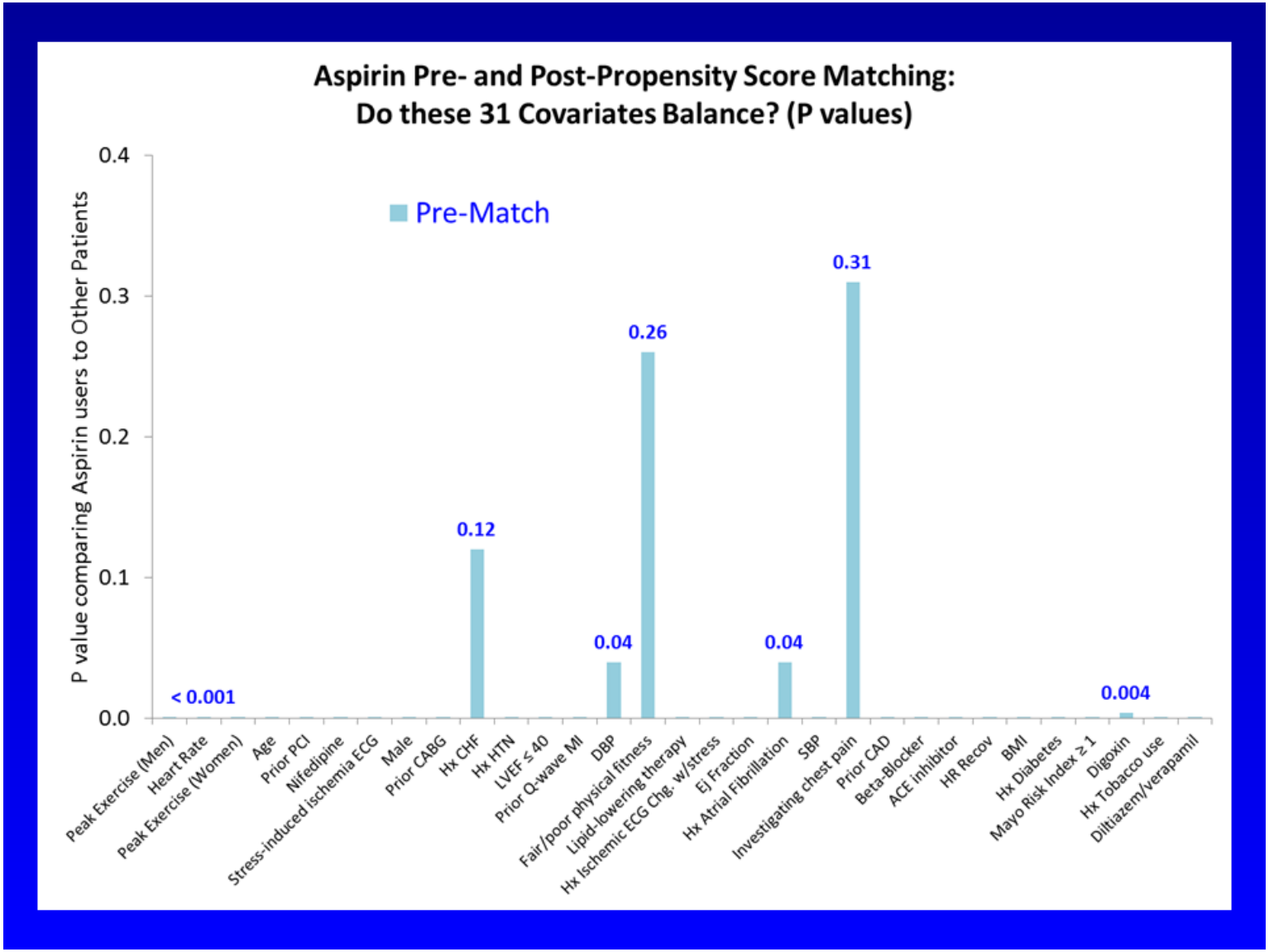
Variable	Aspirin* (n = 1351)	No Aspirin* (n = 1351)	P value
Age, years	60 (11)	61 (11)	.16
Body mass index, kg/m^2	29 (6)	29 (6)	.83
Ejection fraction, %	51 (8)	51 (9)	.65
Resting heart rate, beats/min	77 (13)	76 (14)	.13
Resting systolic BP, mm Hg	141 (21)	141 (21)	.68
Resting diastolic BP, mm Hg	85 (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

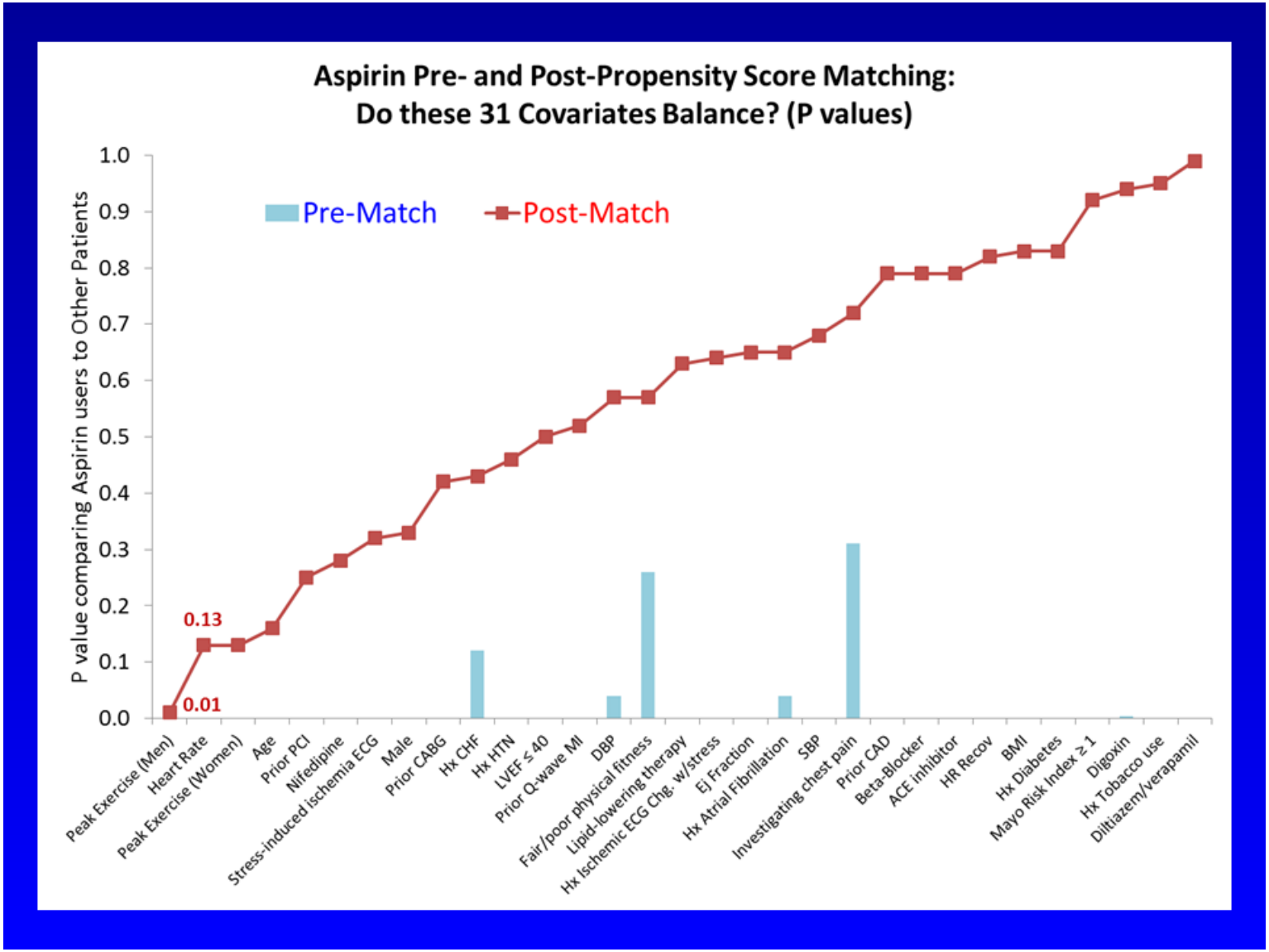
*Cells contain mean (SD)

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

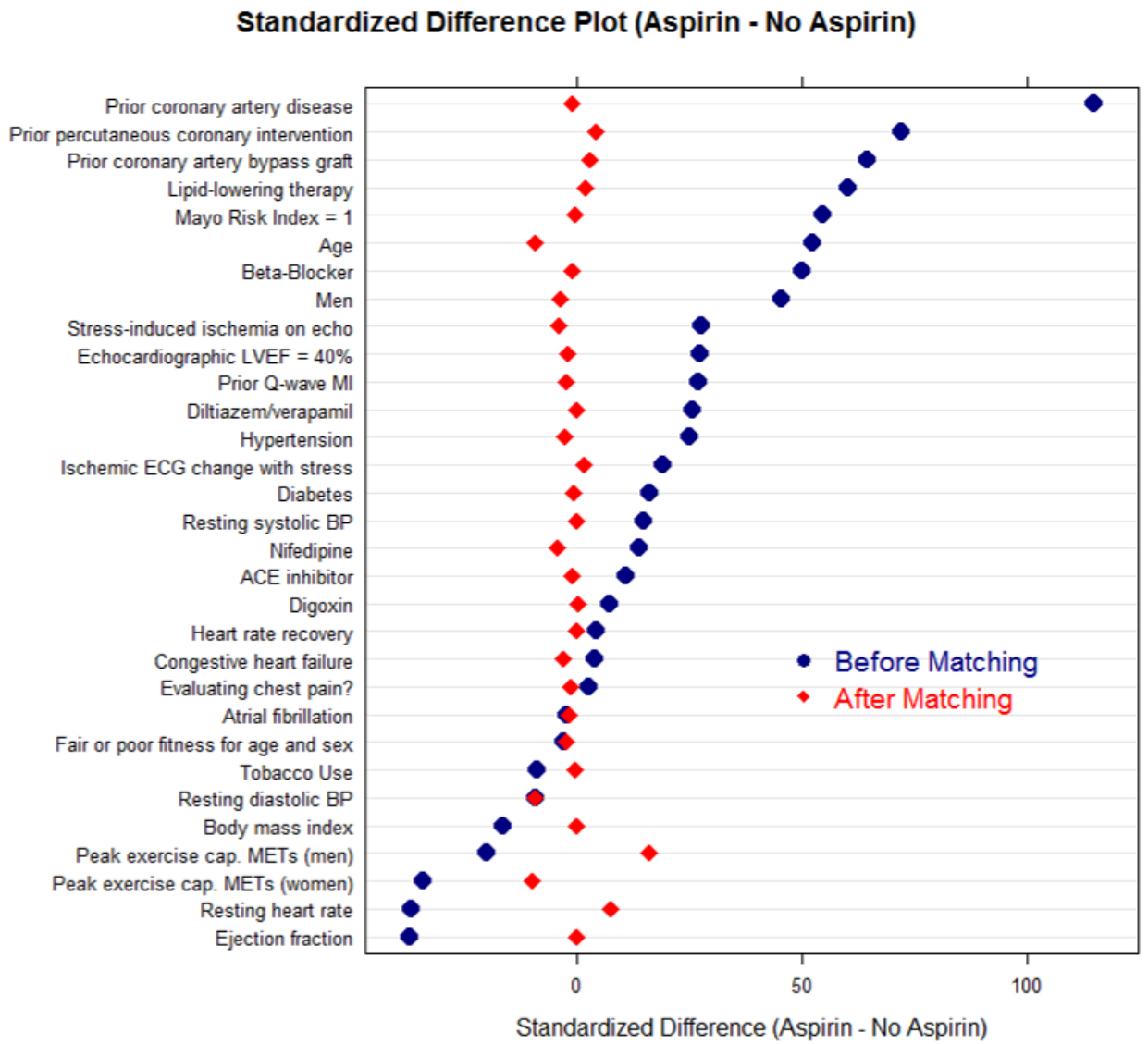
Variable	Aspirin (n = 1351)	No Aspirin (n = 1351)	P value
Men	70.4	72.1	.33
Clinical history: diabetes	15.0	15.3	.83
hypertension	50.3	51.7	.46
prior coronary artery disease	48.3	48.8	.79
congestive heart failure	5.8	6.6	.43
Medication use: Beta-blocker	26.1	26.5	.79
ACE inhibitor	15.5	15.8	.79

- 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

Standardized Difference Plot (Love Plot)

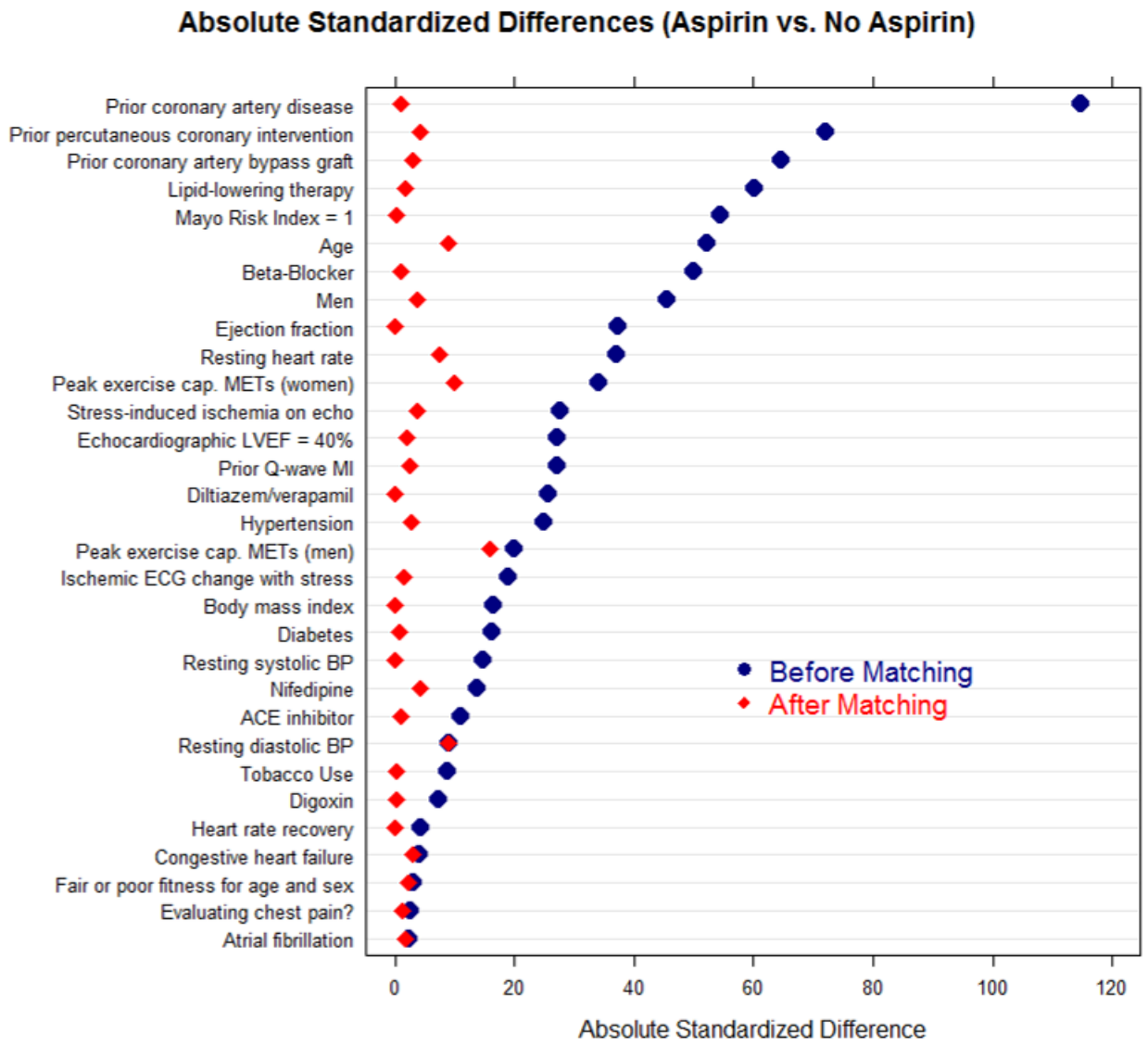


Dotplots of Standardized Differences (Love Plots)

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.

Absolute Standardized Differences Plot



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

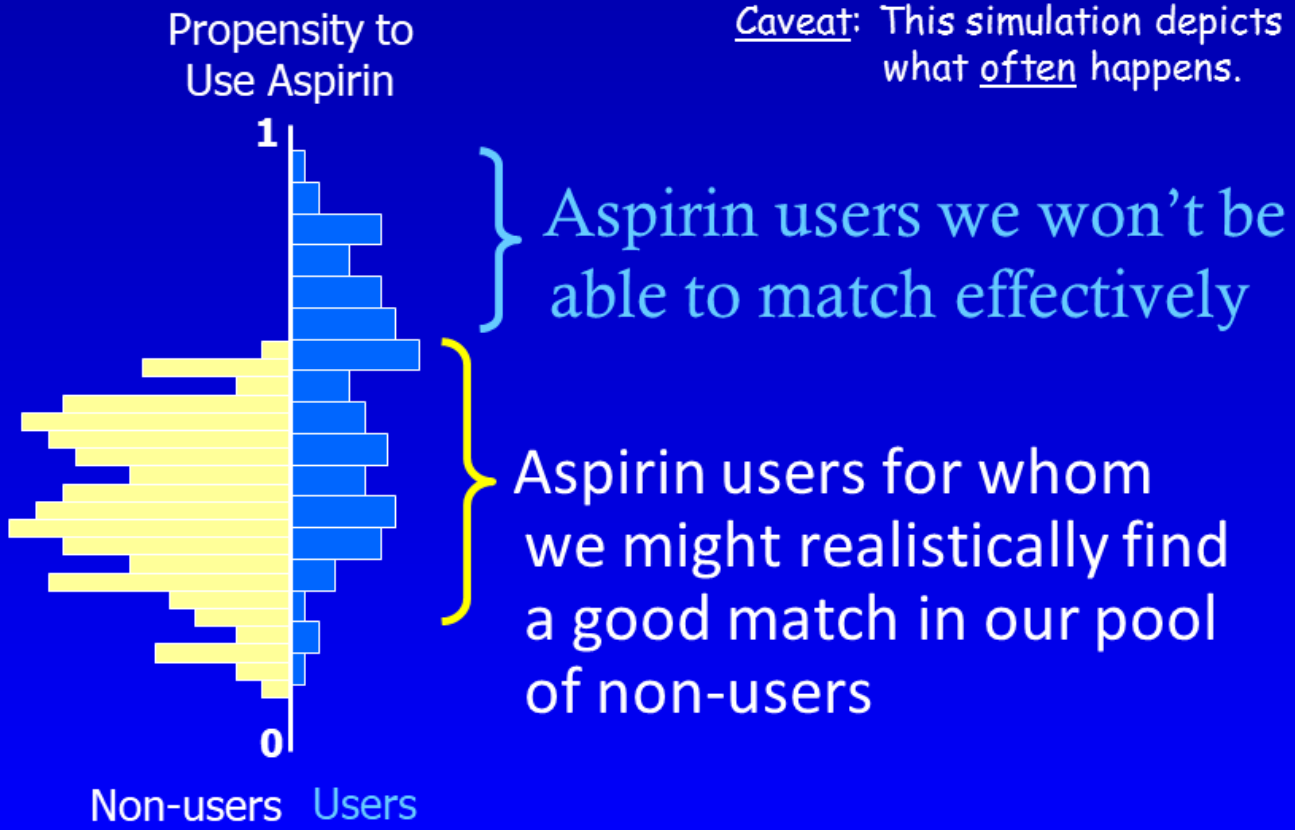
- Trade-off between
  - Failing to match all treated subjects (incomplete)
  - Matching dissimilar subjects (inexact matching)
- Severe bias due to incomplete matching: so that it's usually better to match all treated subjects, then follow with analytical adjustments for residual imbalances in the covariates.
- 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?



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)

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[7]?

- 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.

[7] 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.

Next Time

- Subclassification and Stratification on the Propensity Score
- Direct Adjustment for the Propensity Score
- Weighting using the Propensity Score