Propensity Scores Workshop

github.com/THOMASELOVE/ichps2018

ICHPS: 2018-01-10

Designing and Analyzing Observational Health Policy Studies Using Propensity Scores

All Materials: https://github.com/THOMASELOVE/ichps2018

This workshop describes and demonstrates effective strategies for using propensity score methods to address the potential for selection bias in observational studies comparing the effectiveness of treatments or exposures.

We review the main analytical techniques associated with propensity score methods (focusing on **matching**, **weighting** and **double robust** techniques) and describe key strategic concerns related to effective propensity score estimation, assessment and display of covariate balance, choice of analytic technique, stability and sensitivity analyses, and communicating results effectively.

Wednesday 2018-01-10: 8 - 10 AM in Crystal Ballroom CD

Overview

All Materials: https://github.com/THOMASELOVE/ichps2018

1. Fundamentals
   - Exposure Selection Bias in an Observational Study
   - What is a propensity score and how should we build one?
2. Using the propensity score to deal with selection bias
   - Via matching
   - Via weighting
   - Double robust approach (weighting + regression)
3. Strategies for effective design and analyses of health policy studies
   - Assessing and Displaying Covariate Balance
   - Stability and Sensitivity Analysis

Fundamentals

Designing A Study: ASA and Mortality in Heart Subjects

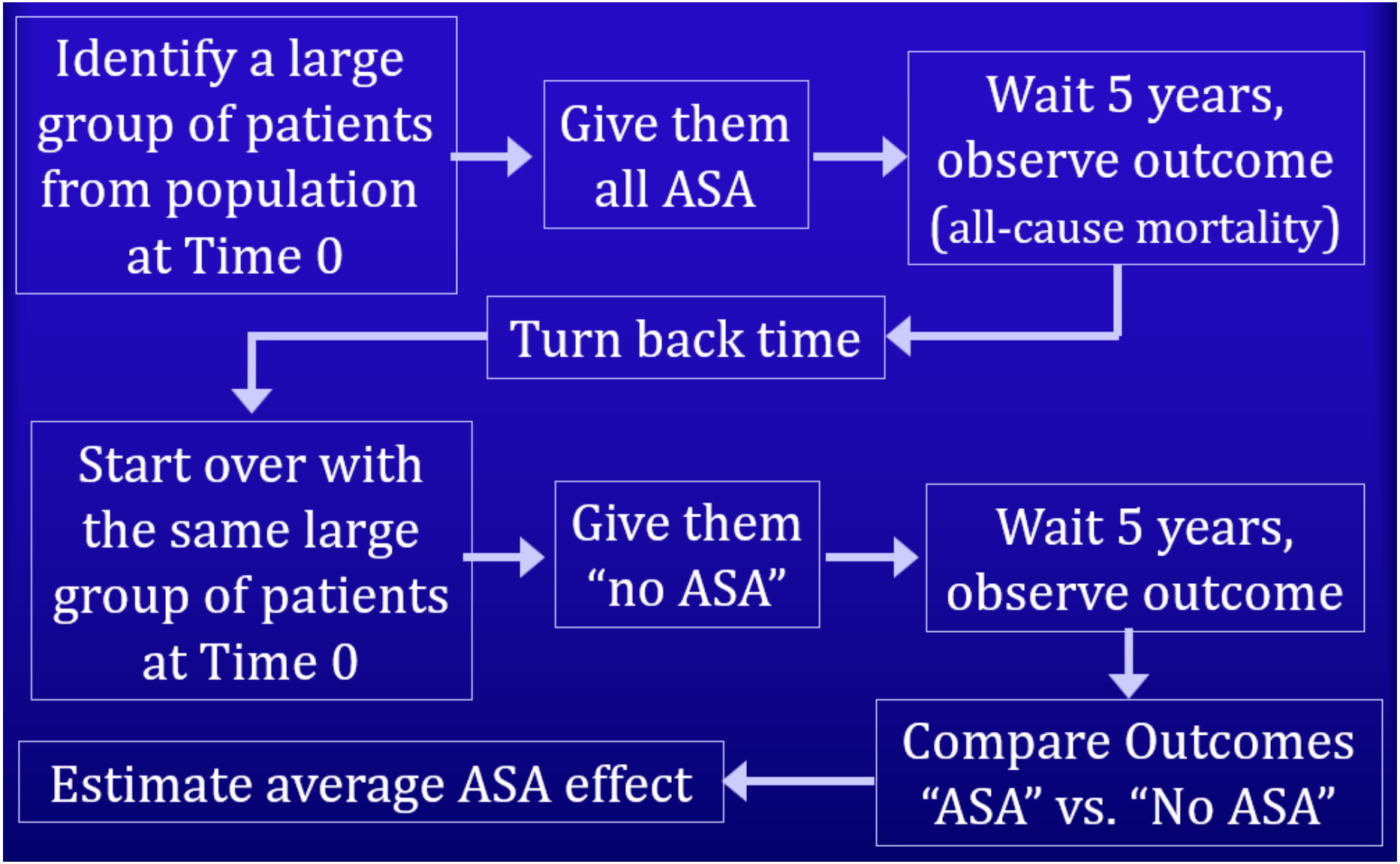
- Suppose you want to study the effect of aspirin (acetylsalicylic acid: ASA) on all-cause mortality.
- You identify an interesting group of Subjects as those undergoing stress echocardiography.
    - Your goal is to compare ASA Subjects to "no ASA" Subjects

What would be the **ideal** study?

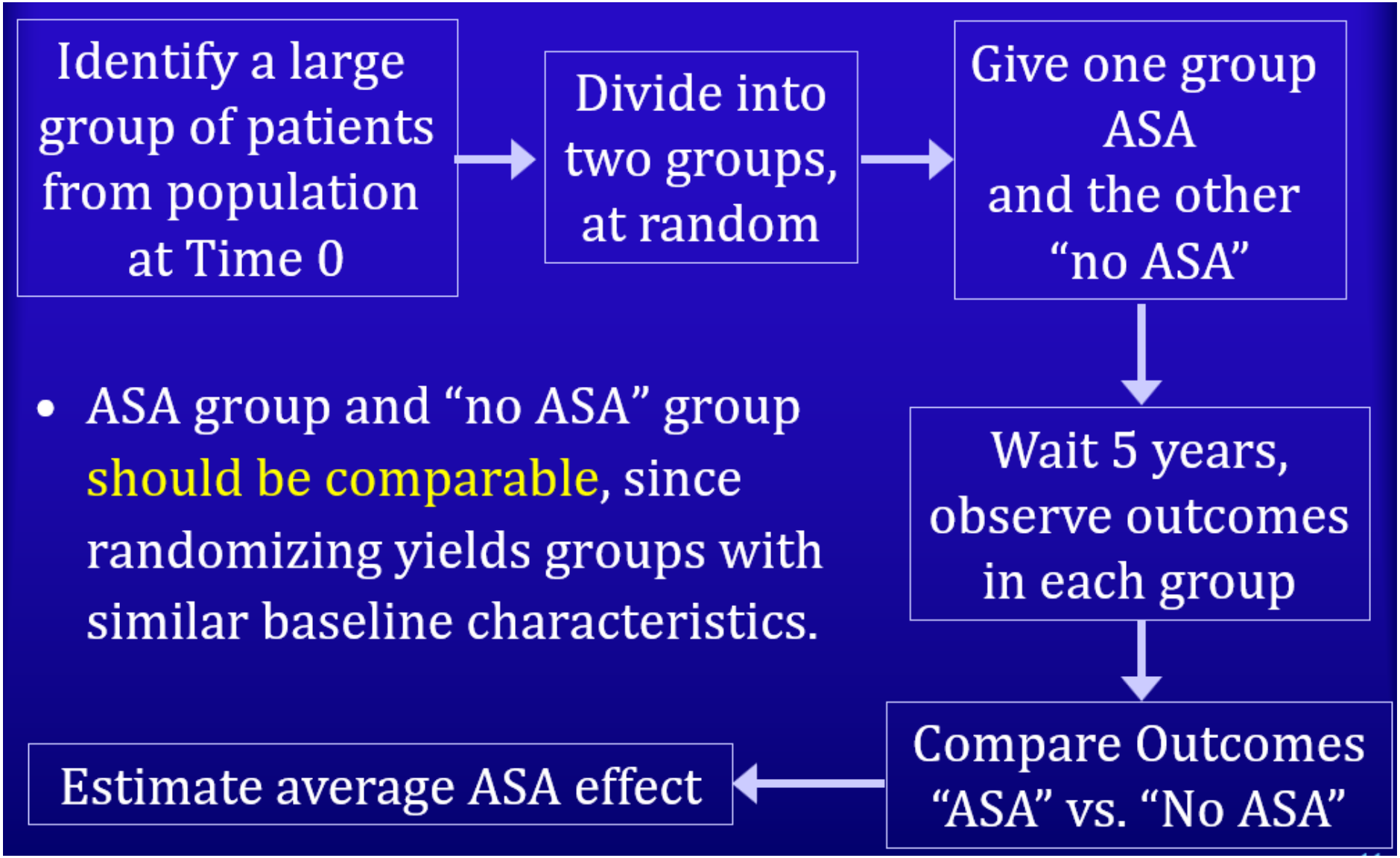
Step 1. Identify a large group of Subjects from the population at Time 0.

Step 2?

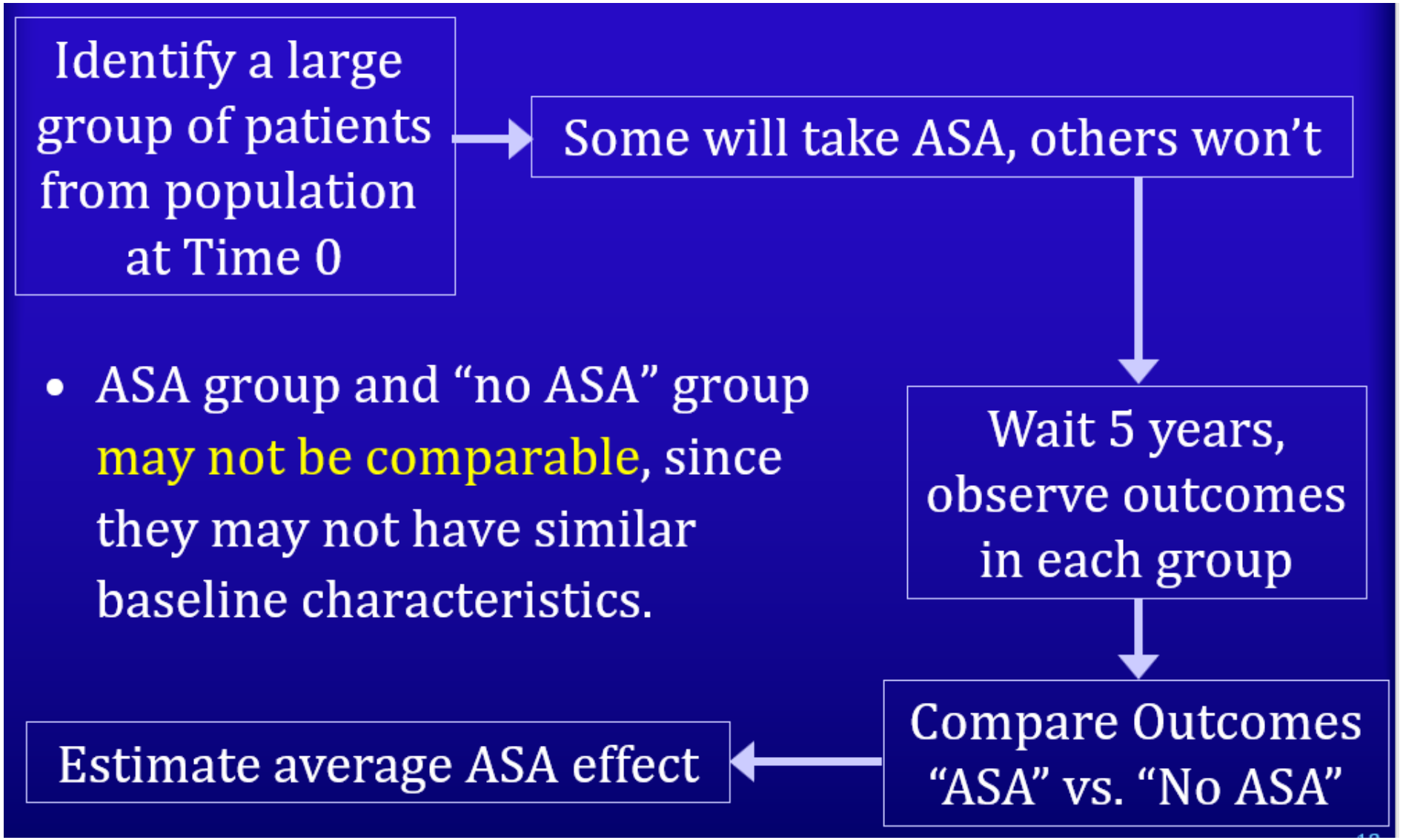
ASA and Mortality: Ideal Study



ASA and Mortality: Best Practical Study?



ASA and Mortality: Observational Study?



Simple Observational Studies

Specify a population, outcome, exposure and covariates.

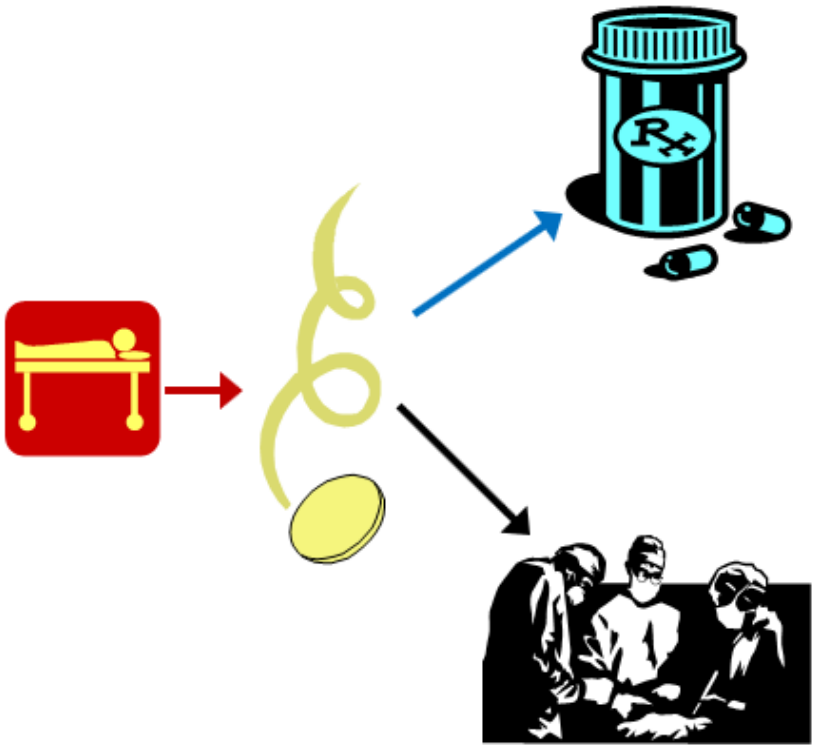
- We want to make a fair comparison between the exposed group and the control group within a population, in terms of an outcome of interest.
- We can obtain covariates that describe the subjects before they were allocated to the exposed and control groups.

But

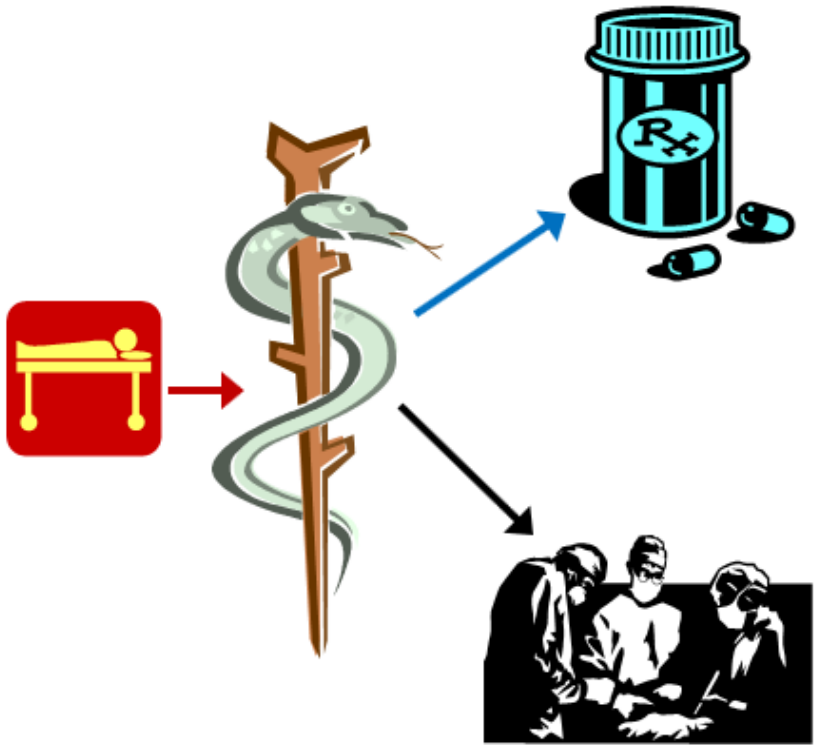
- We **cannot** use randomization to ensure that the groups will be comparable in terms of the covariates.

No randomization forces the investigator to think hard about how the exposures are "assigned"...

Randomized Trials vs. Observational Studies



Randomization "ensures" that subjects receiving different exposures are comparable.



In observational studies, the researcher does not randomly allocate the exposures.

How Do We Avoid Being Misled by Observational Studies?

- What differentiates an observational study from a randomized controlled trial?
  - One key element: potential for selection bias.
- What is selection bias and what can we do about it?
  - Baseline characteristics of comparison groups are different in ways that affect the outcome.

We will often distinguish between overt and hidden bias.

- Overt Bias (seen in data - propensity scores can help)
- Hidden Bias (required data not collected - requires sensitivity analyses)

What do you want to know about an intervention?

- 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?

This is based on a marketing list originally posted at anabus.com

The database you wish you had...

Subject	Outcome if Subject receives Aspirin	Outcome if Subject does not receive Aspirin	Aspirin Effect
A	10	6	+4
B	9	5	+4
C	7	3	+4
D	11	8	+3
E	6	3	+3
F	9	6	+3

Harsh Reality

Subject	Outcome if Subject receives Aspirin	Outcome if Subject does not receive Aspirin	Aspirin Effect
A	10	-	?
B	9	-	?
C	7		?
D	-	8	?
E	-	3	?
F	-	6	?

This is based on a marketing list originally posted at anabus.com

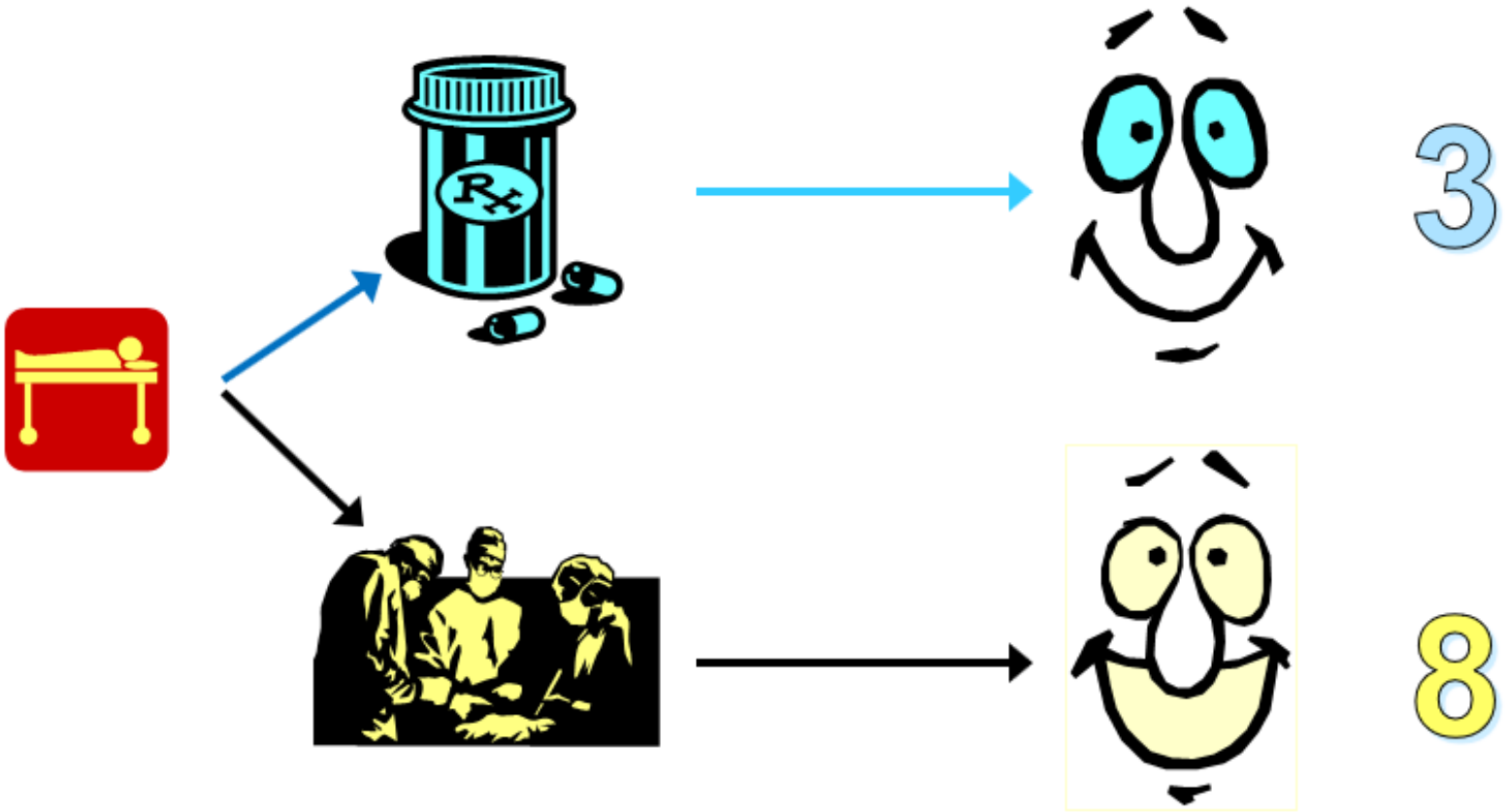
Rubin Causal Model

Builds on the potential outcomes framework[1].

- Key assumption: Strongly Ignorable Exposure Assignment
  - Potential outcomes Y_{ASA}, Y_{no_ASA} are assumed conditionally independent of exposure assignment, given covariates $\mathbf{X}$.
  - This is a "no hidden bias" assumption, where we assume $\mathbf{X}$ contains all relevant information about exposure assignment.
- Data? Clinical-Researchers vs. Economists

[1] Rubin 1997, Rosenbaum 2002

Causal Effects in terms of "Potential Outcomes"



Surgery - Drug Effect = - = 8 - 3

Assessing the Causal Effect of an Exposure on an Outcome

- Objective: Draw causal inferences between [exposed vs. control] and outcome
- Standard Approach: Risk Adjustment
  - Problem: Selection Bias (people getting exposure are different from people getting control in ways that affect outcome)
- Idea: Compare exposed to control subjects that looked similar (had similar propensity for exposure) prior to the exposure decision

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[2]

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

[2] McCaffrey et al 2004 describe boosted regression approaches.

The New Database, Simply

Subject	Propensity to receive Aspirin	Outcome if Subject receives Aspirin	Outcome if Subject does not receive	Aspirin Effect
A	0.81	10	-	?
B	0.51	9	-	?
C	0.31	7		?
D	0.79	-	8	?
E	0.51	-	3	?
F	0.29	-	6	?

We could then match up the subjects, and plug in the estimates, for instance.

- So how do we model propensity for aspirin, or any other exposure?

Our Goal

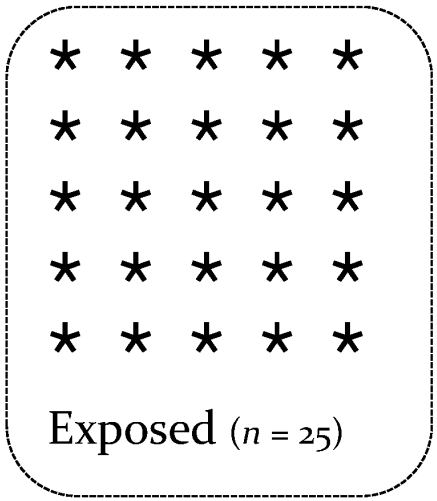
Fit a model that predicts the probability of exposure, given a set of covariates.

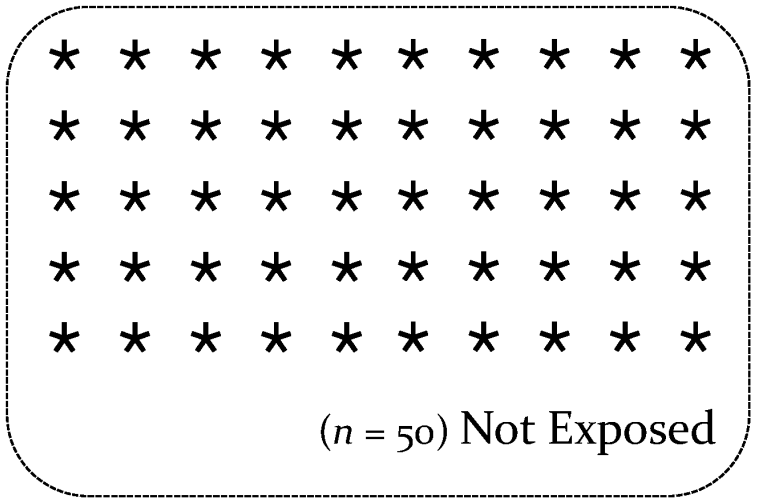
We anticipate that our exposure group will have different distributions of the covariates than our control group at baseline.

- Perhaps the subjects taking aspirin are older
- Perhaps they are more likely to be taking other medications
- Perhaps they are more likely to have certain comorbidities

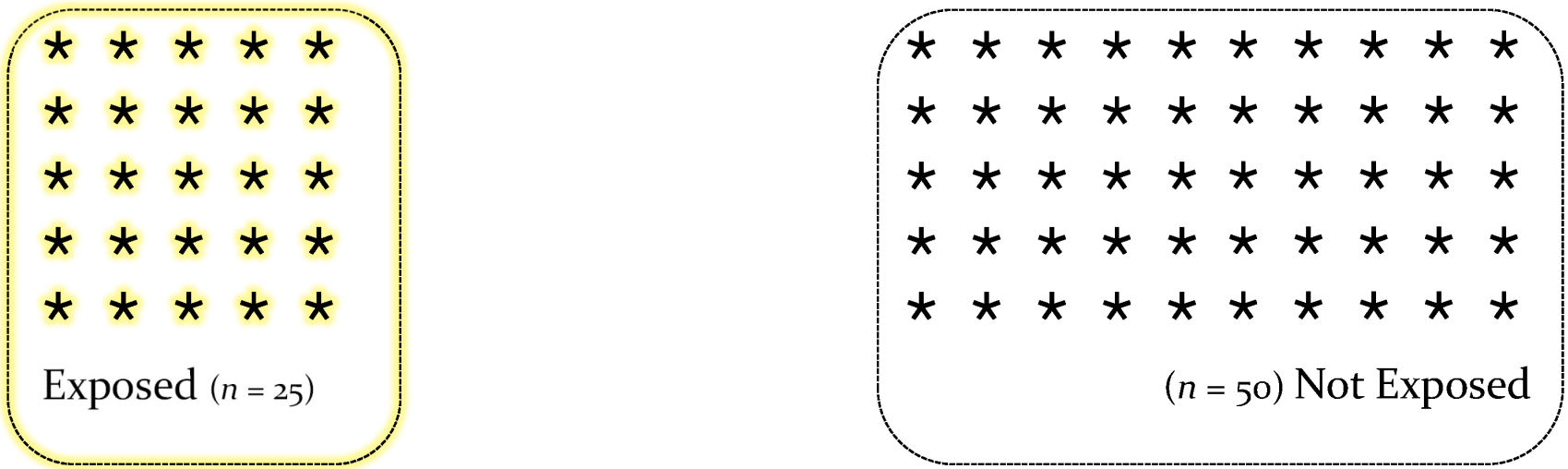
We want to wind up with a fair basis for comparison between exposed and control subjects.

Simple Observational Study

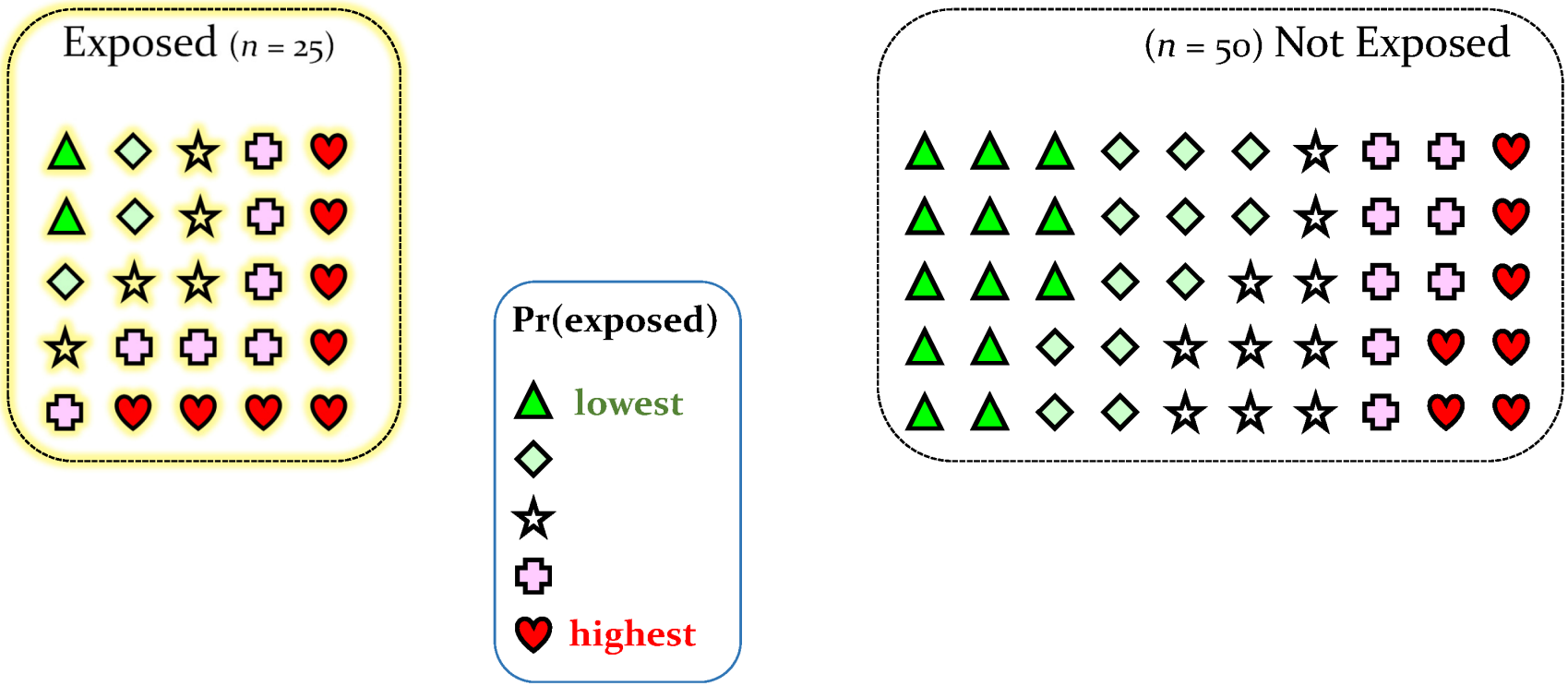




Apply the Exposure



Characterize by propensity to receive the exposure...



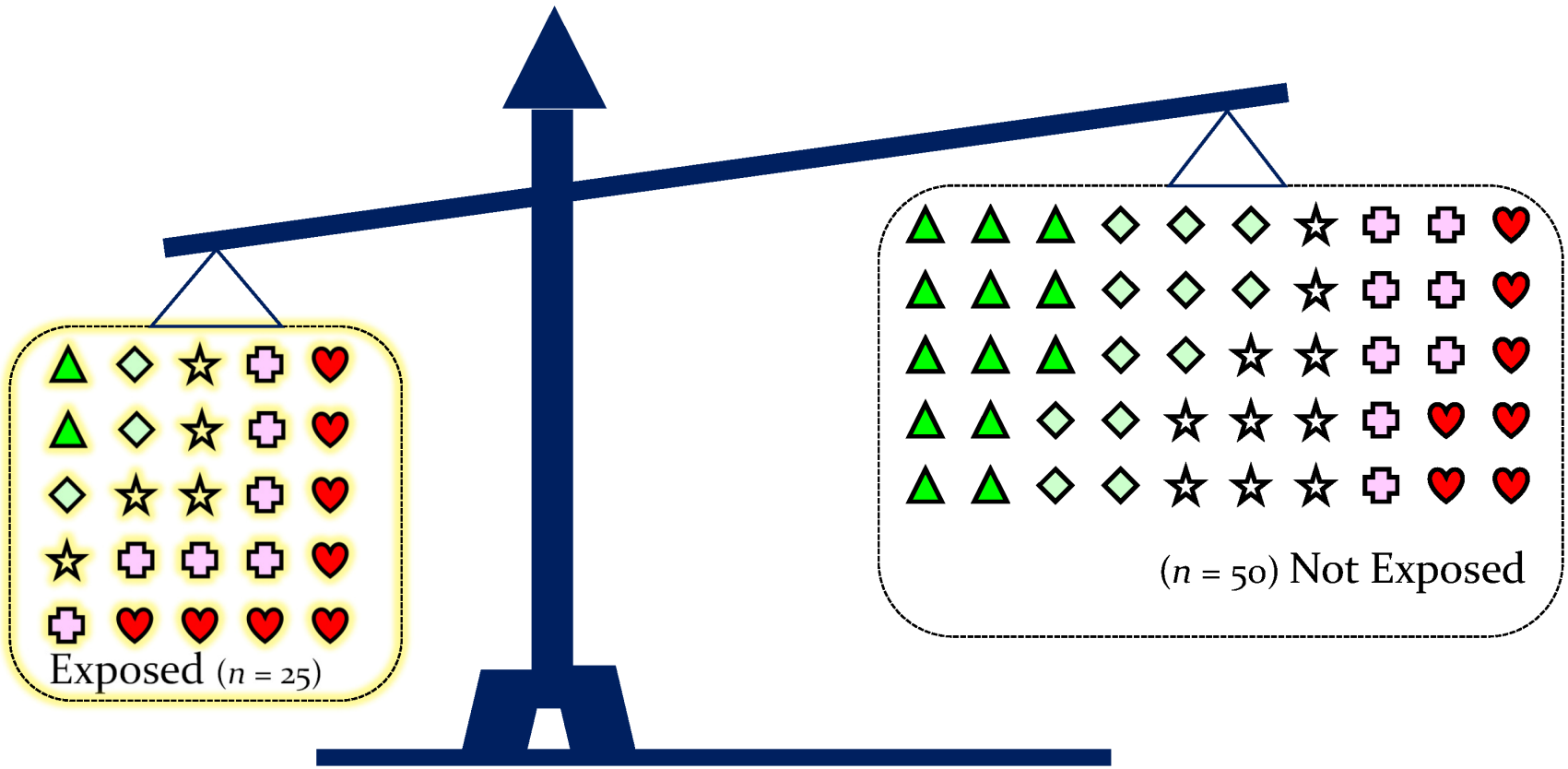
Estimating Propensity for Exposure Given Covariates

Usual Tool: Logistic Regression Model for Exposure Allocation

- Consider including any variables that have a relationship to the exposure decision
  - Precede the exposure in time
  - Relevant to exposure assignment
  - Relationship to outcome? (some controversy here)
- No information included on the actual exposure received, or on the outcome(s)
- In early stages, always err on the side of inclusion

$$\text{logit}[\Pr(\text{aspirin})] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Statin} + ... + \beta_k \text{Diabetes}$$

Are baseline characteristics in balance?

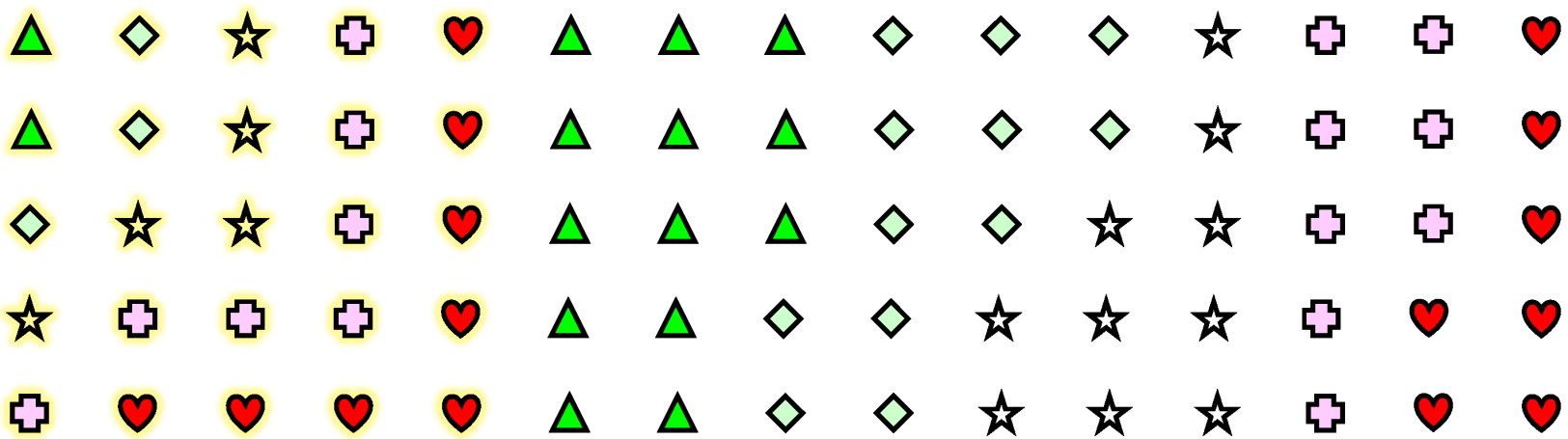


Model Without the Propensity Score

Outcome = β_0 + β_1*Exposure,

for pool of 75 subjects

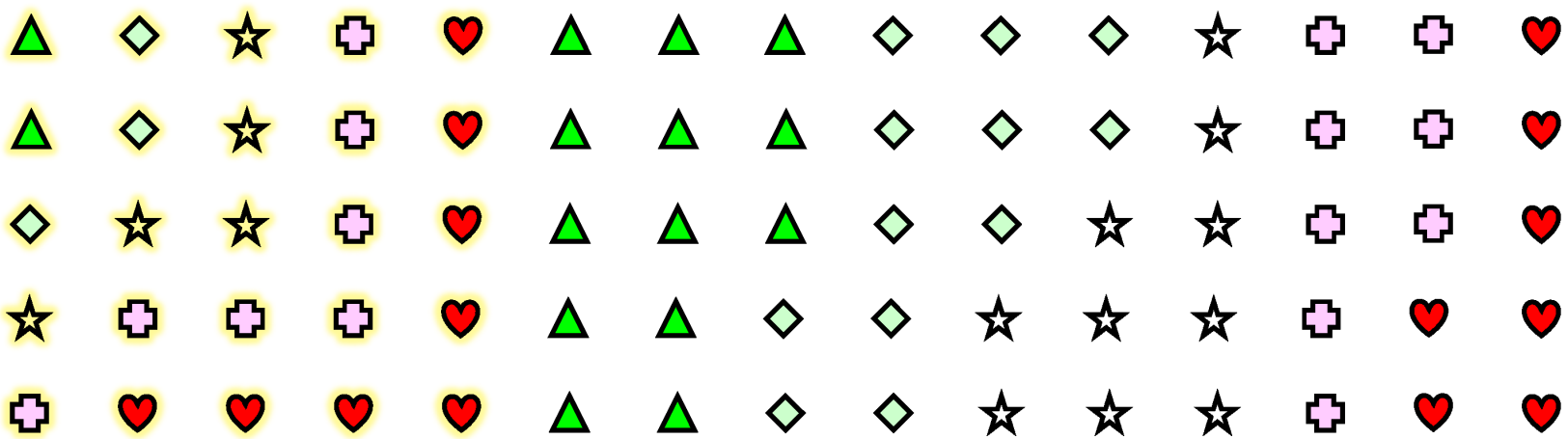
We interpret β_1 as the exposure's effect.



Direct Adjustment for Propensity Score

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

Again, across entire pool of 75 subjects



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[3]?
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?

[3] 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.

Matching Using The Propensity Score

Multivariate Matching using the Propensity Score

Goal: Emulate a randomized clinical trial in matching[4], then use standard analyses to compare matched sets.

Design: Exposed subjects matched to people who didn't receive exposure but who had similar propensity to receive exposure (matching exposed to unexposed "clones")

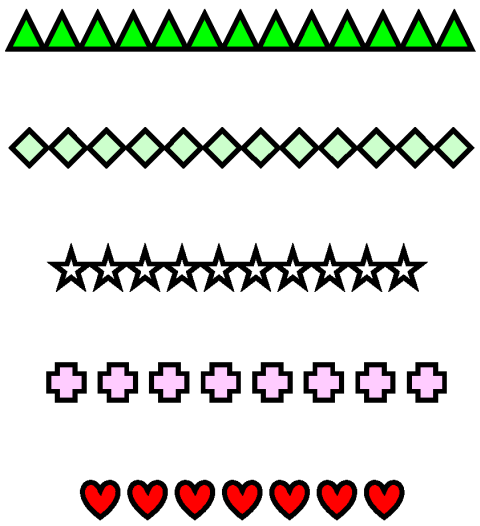
- Match subjects on a scalar (the propensity score) so that the **groups** are similar in terms of distributions of multiple covariates simultaneously.

[4]Seminal paper: Rosenbaum & Rubin 1985. See also Rosenbaum 2010 and Rosenbaum 2017

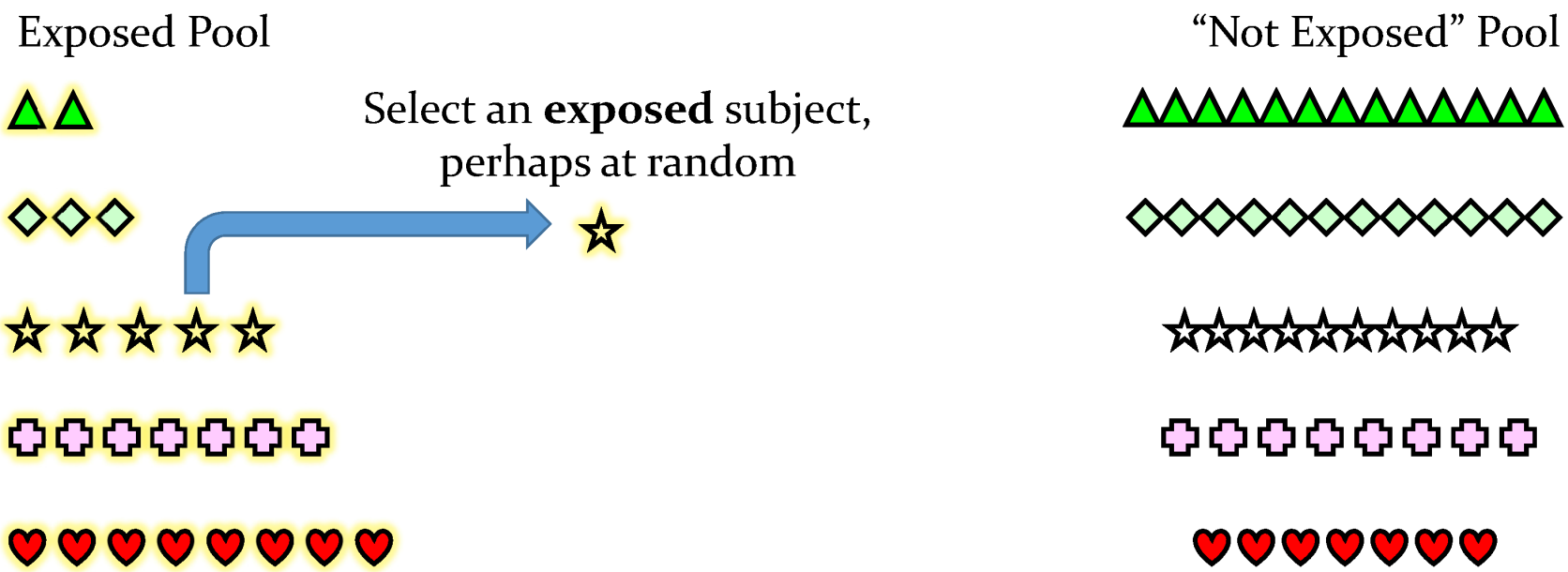
Propensity Score Matching (1:1)

Exposed Pool

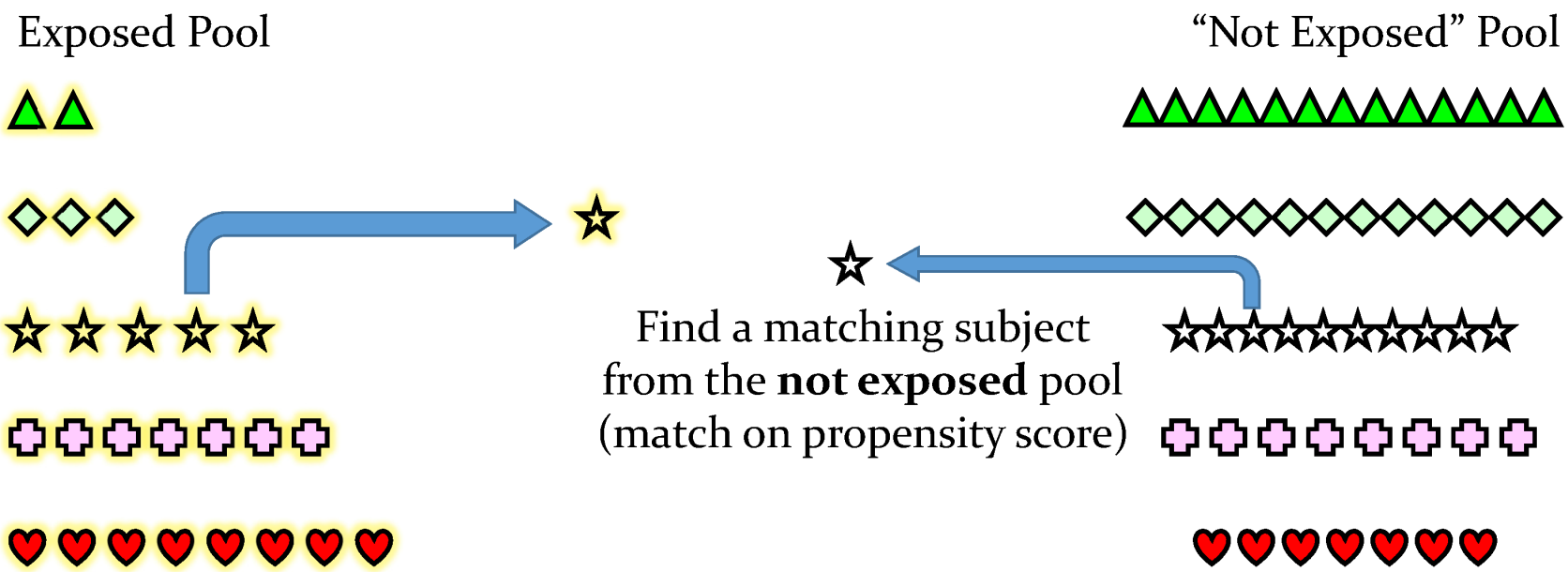
"Not Exposed" Pool



Propensity Score Matching (1:1)



Propensity Score Matching (1:1)



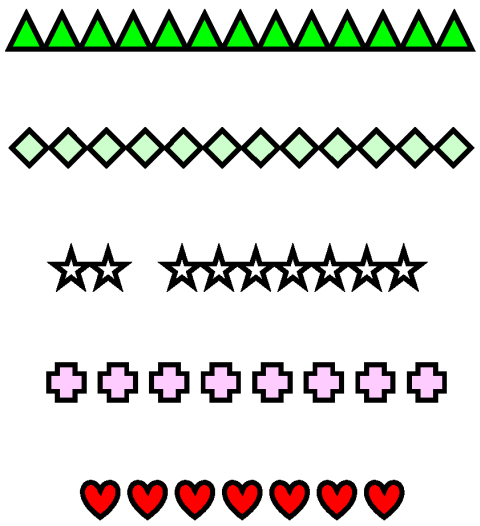
Propensity Score Matching (1:1)

Exposed Pool

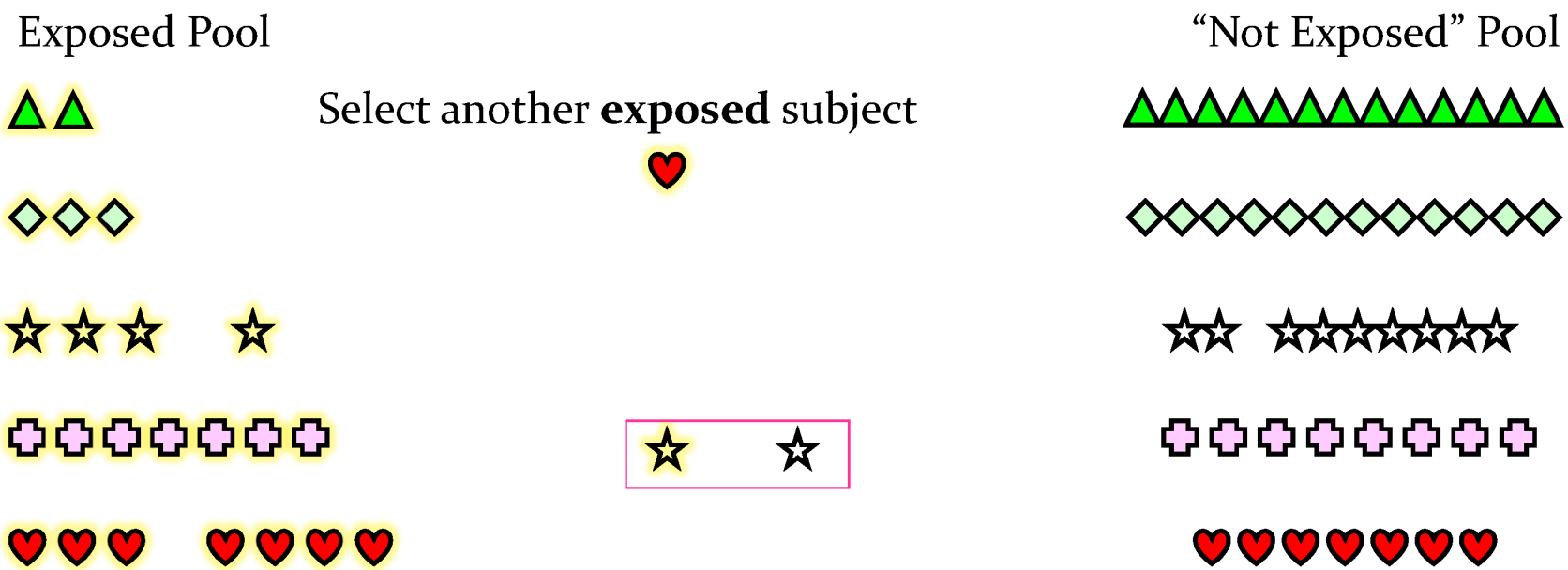
Form a matched pair

We're matching without replacement.

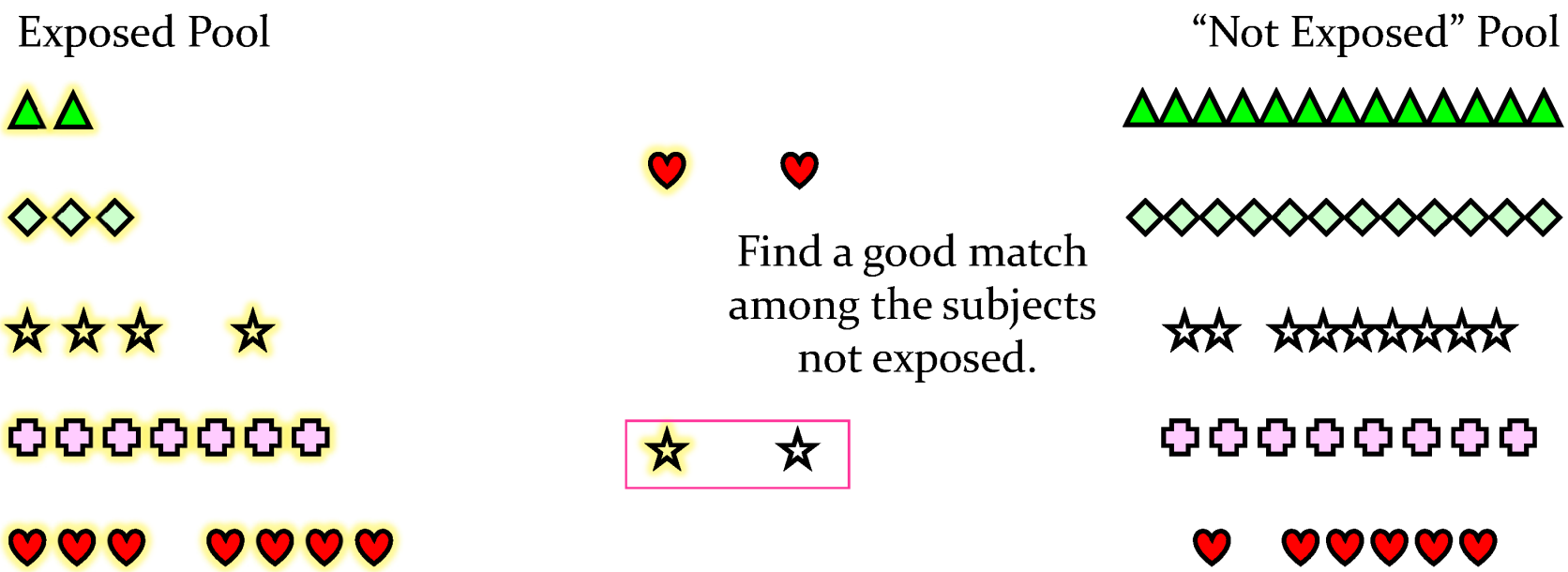
"Not Exposed" Pool



Propensity Score Matching (1:1)



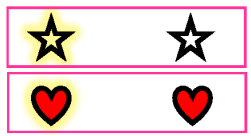
Propensity Score Matching (1:1)



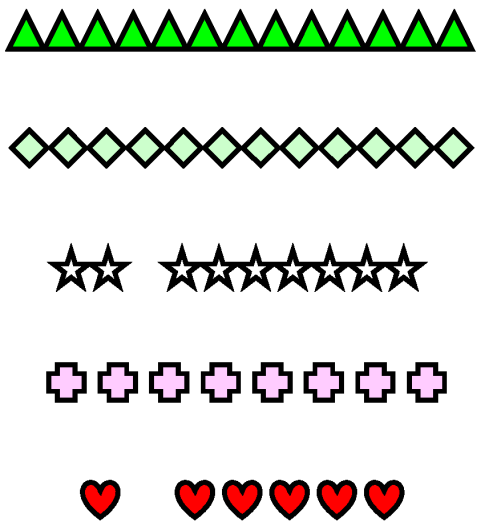
Propensity Score Matching (1:1)

Exposed Pool

A second matched pair!



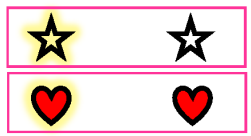
"Not Exposed" Pool



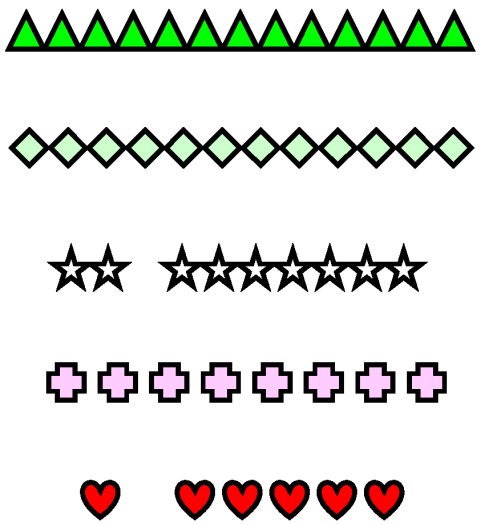
Propensity Score Matching (1:1)

Exposed Pool

Keep matching, until we can find no more acceptable matches



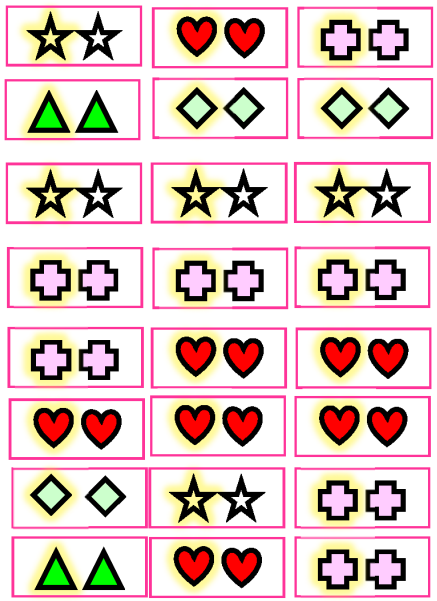
"Not Exposed" Pool



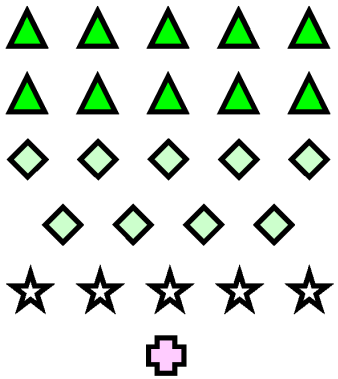
Propensity Score Matching (1:1)

Exposed Pool
(unmatched)

Matched Set
(24 pairs)

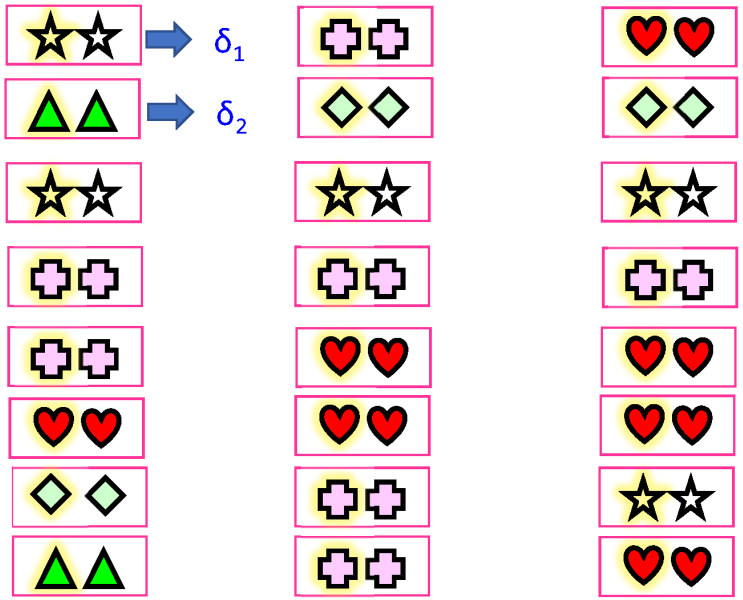


"Not Exposed" Pool
(unmatched)



Propensity Score Matching (1:1)

Matched Set
(24 pairs)

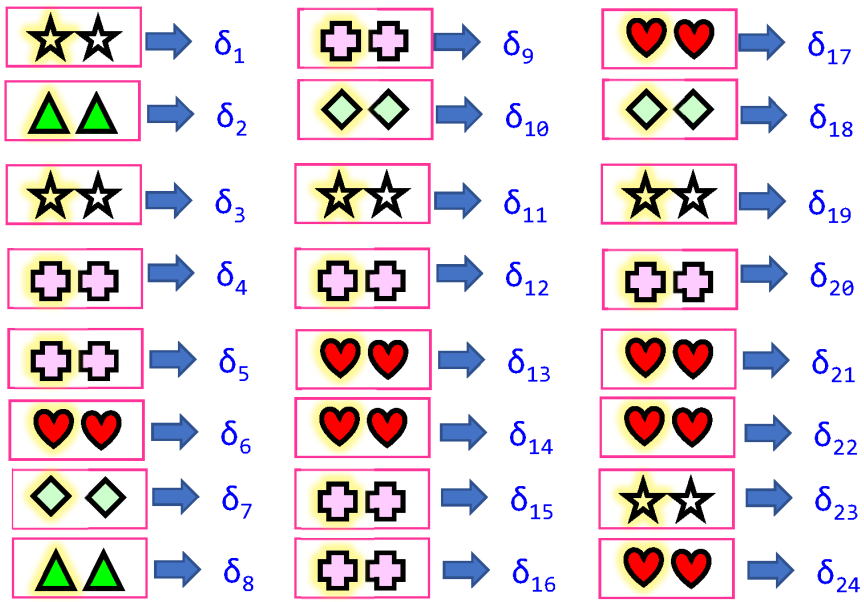


Within each matched pair, compare outcome in exposed subject to outcome in "not exposed" subject.

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

The Aspirin Use and Mortality Study

Gum et al. 2001 JAMA

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

Study of 6,174 adults at Cleveland Clinic (1990-1998) undergoing stress echocardiography to evaluate coronary disease[5].

- 2,310 (37%) were taking aspirin (the exposure)
- 31 covariates are reported, including demographics, clinical history, medications, cardiovascular assessments, and exercise capacity
- 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).

[5] see Gum et al. 2001. The study began with 9,954 consecutive adults.

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?

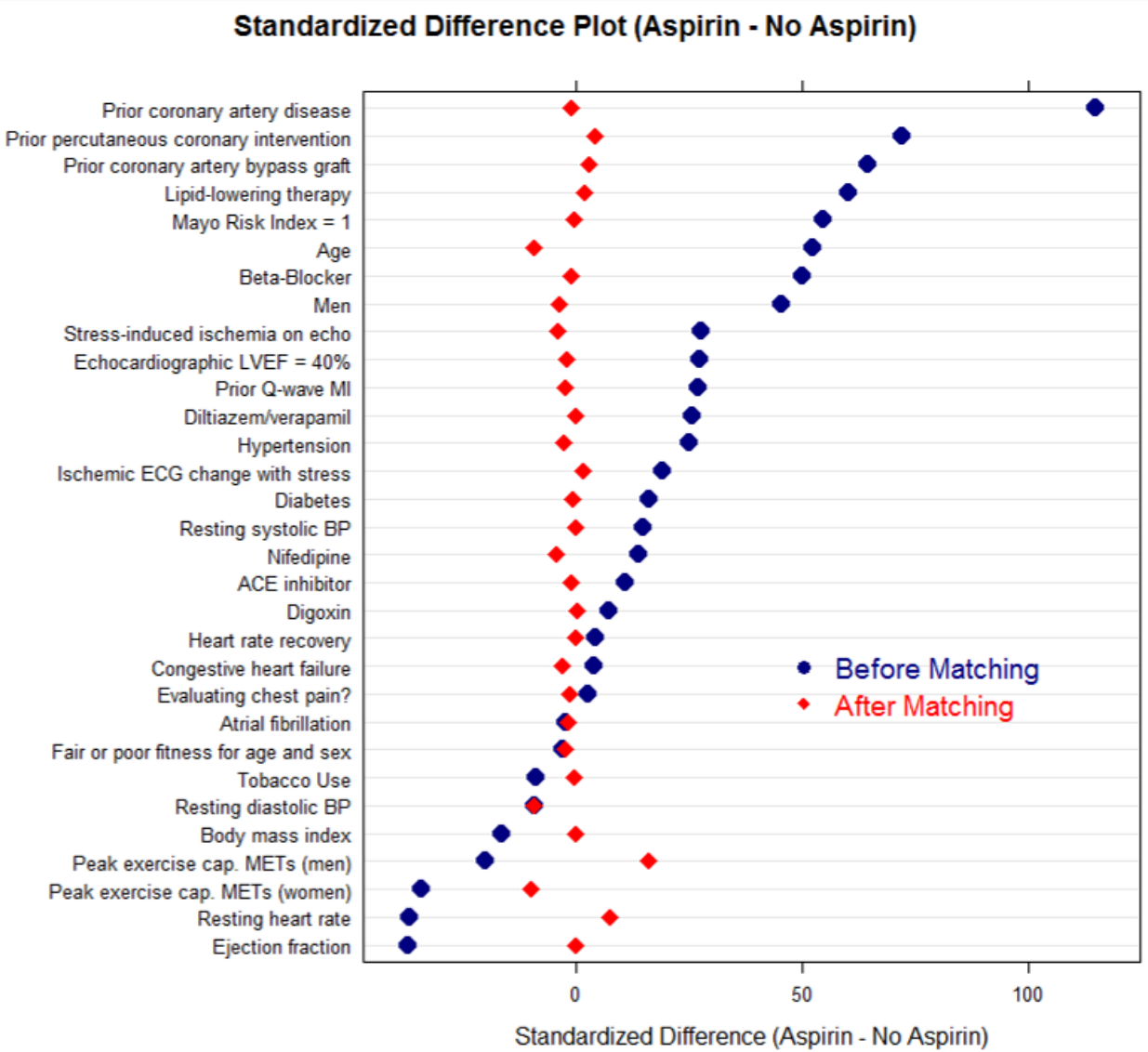
- At baseline, aspirin patients display higher risk of mortality, in general
  - they are older, more likely to be male, and more likely to have a clinical history
  - they are more likely to be on other medications than non-aspirin subjects
  - their cardiovascular assessments are (generally) worse 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%

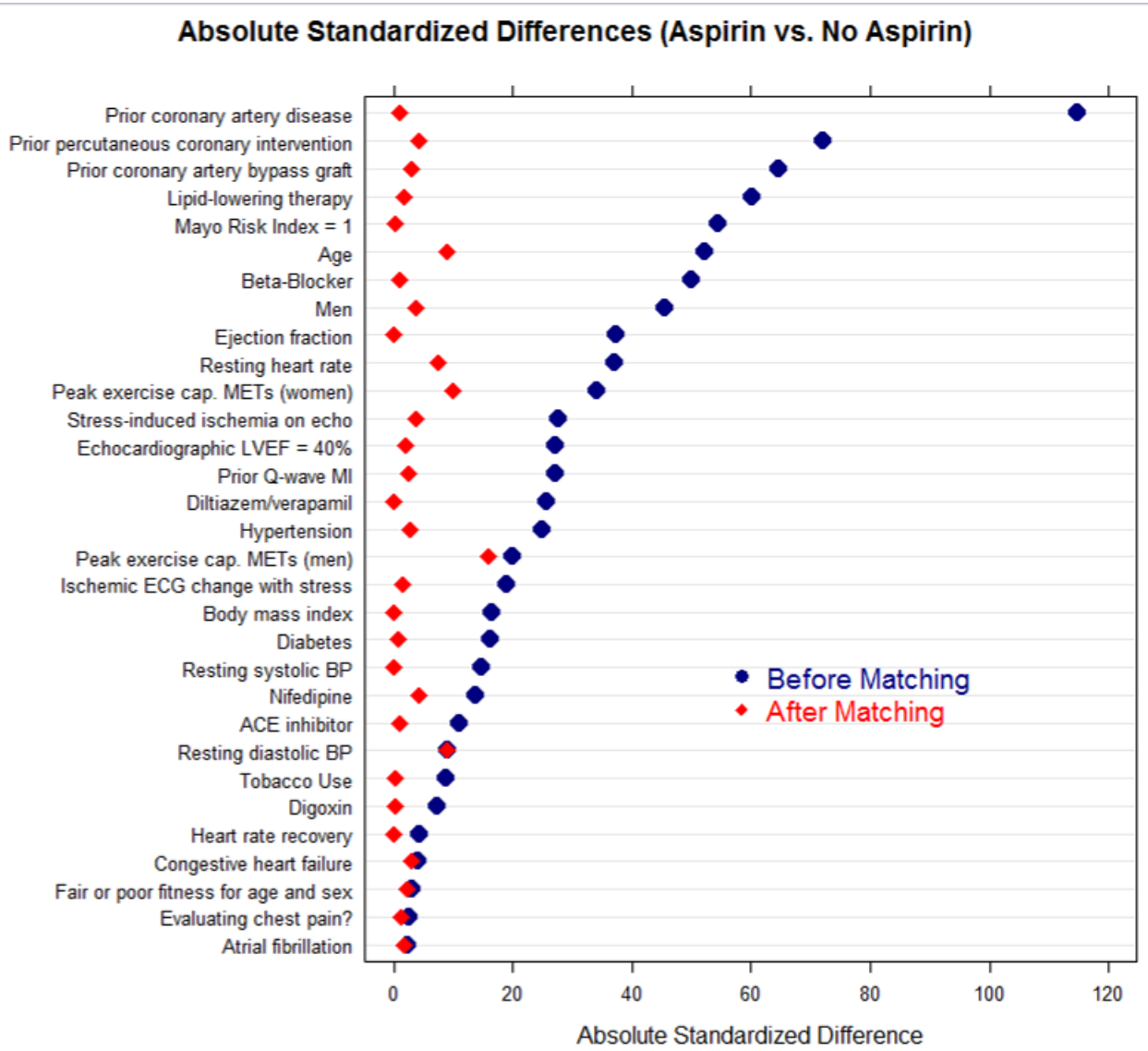
Propensity Score Matching

A logistic regression model was used to estimate the propensity for aspirin use (31 covariates, main effects) for each of the 6,174 subjects. C = 0.83.

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.


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)


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

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[6]?

- 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.

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

Matching

Matching is a fundamental part of the toolbox[7].

- 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.
- 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

Weighting using the Propensity Score

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

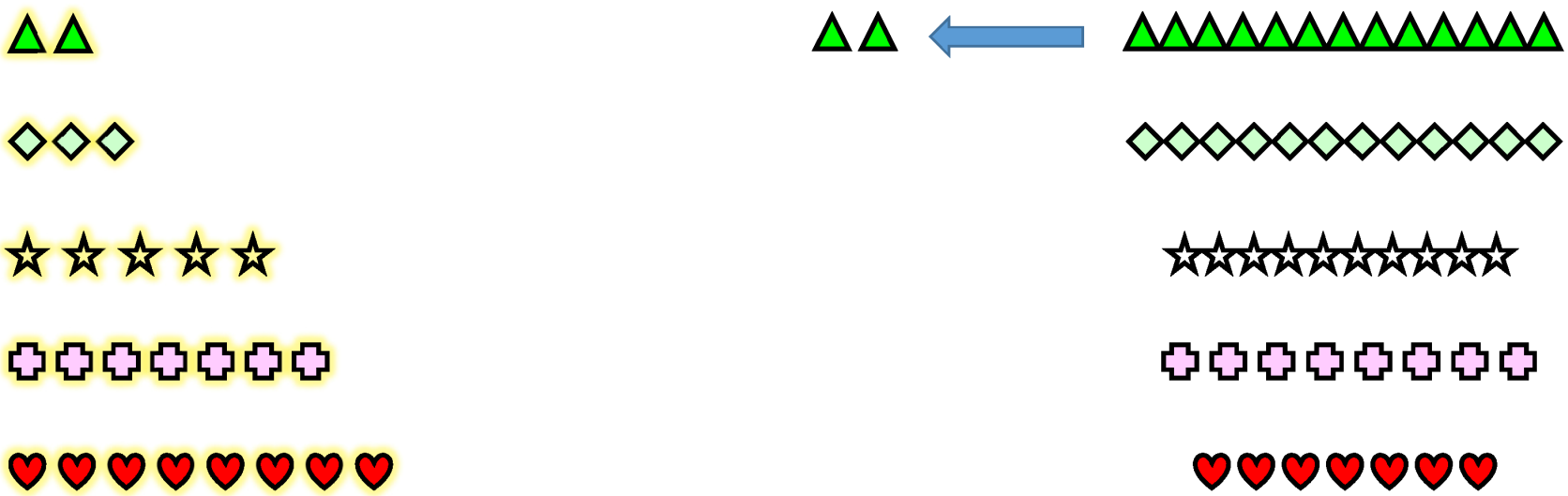
"Not Exposed"
unweighted

Propensity Score Weighting ("ATT")

All Exposed
get weight 1

"Not Exposed"
weighted

"Not Exposed"
unweighted

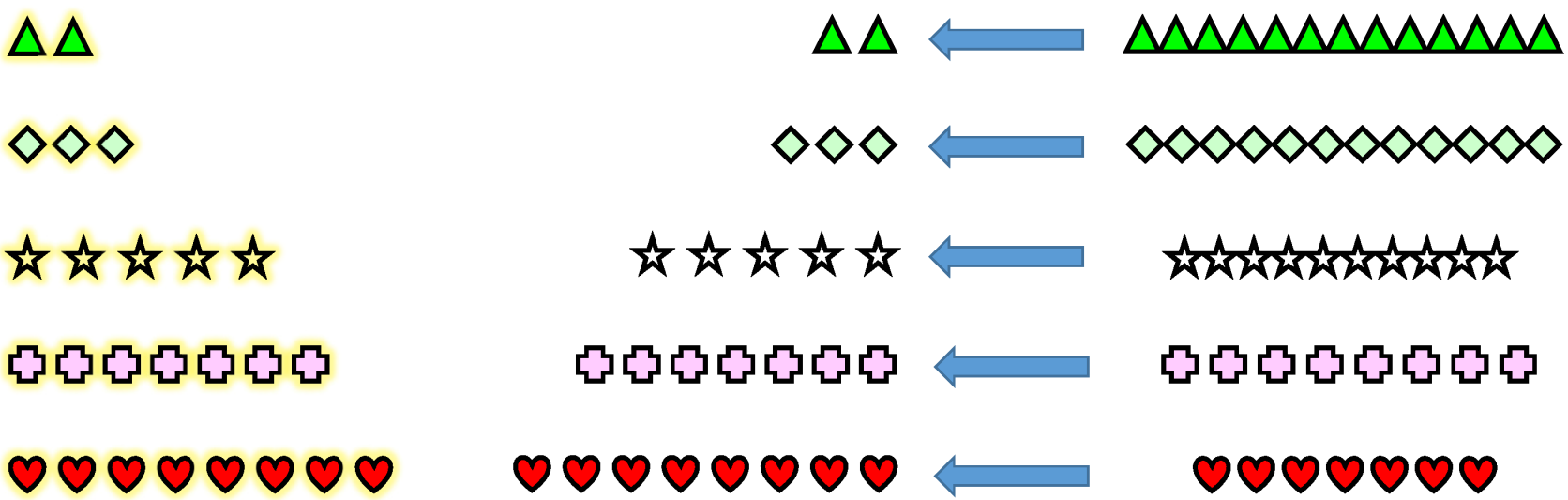


Propensity Score Weighting ("ATT")

All Exposed get weight 1

"Not Exposed" weighted

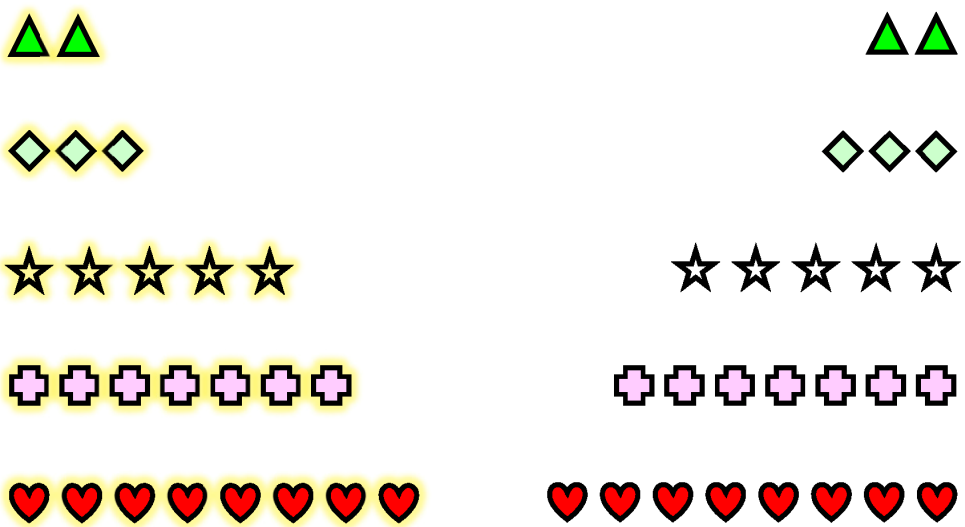
"Not Exposed" unweighted



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[8].

- 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.

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

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

The Support / RHC Study

Studying Right Heart Catheterization in SUPPORT

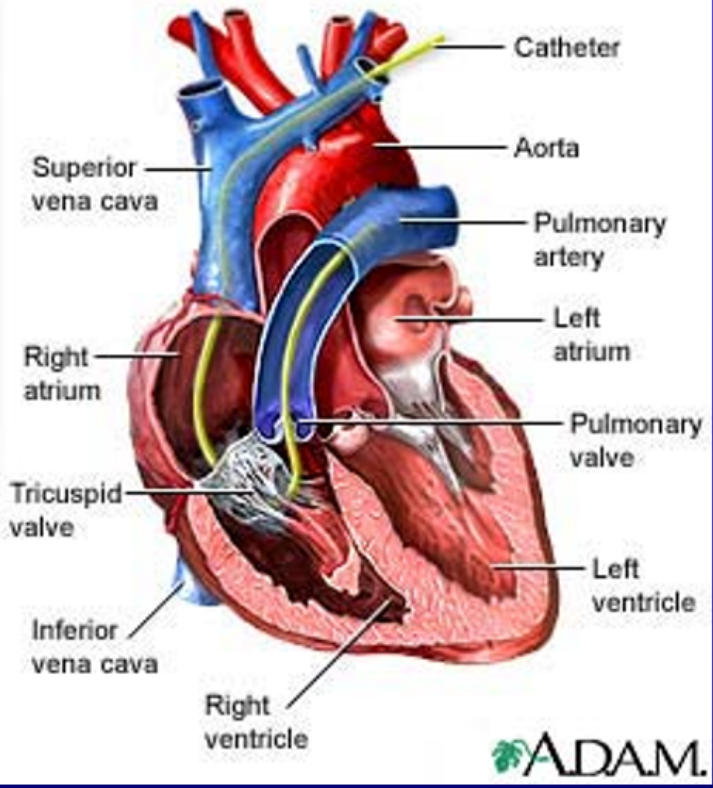
SUPPORT: Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments[9]

- 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!

[9] 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

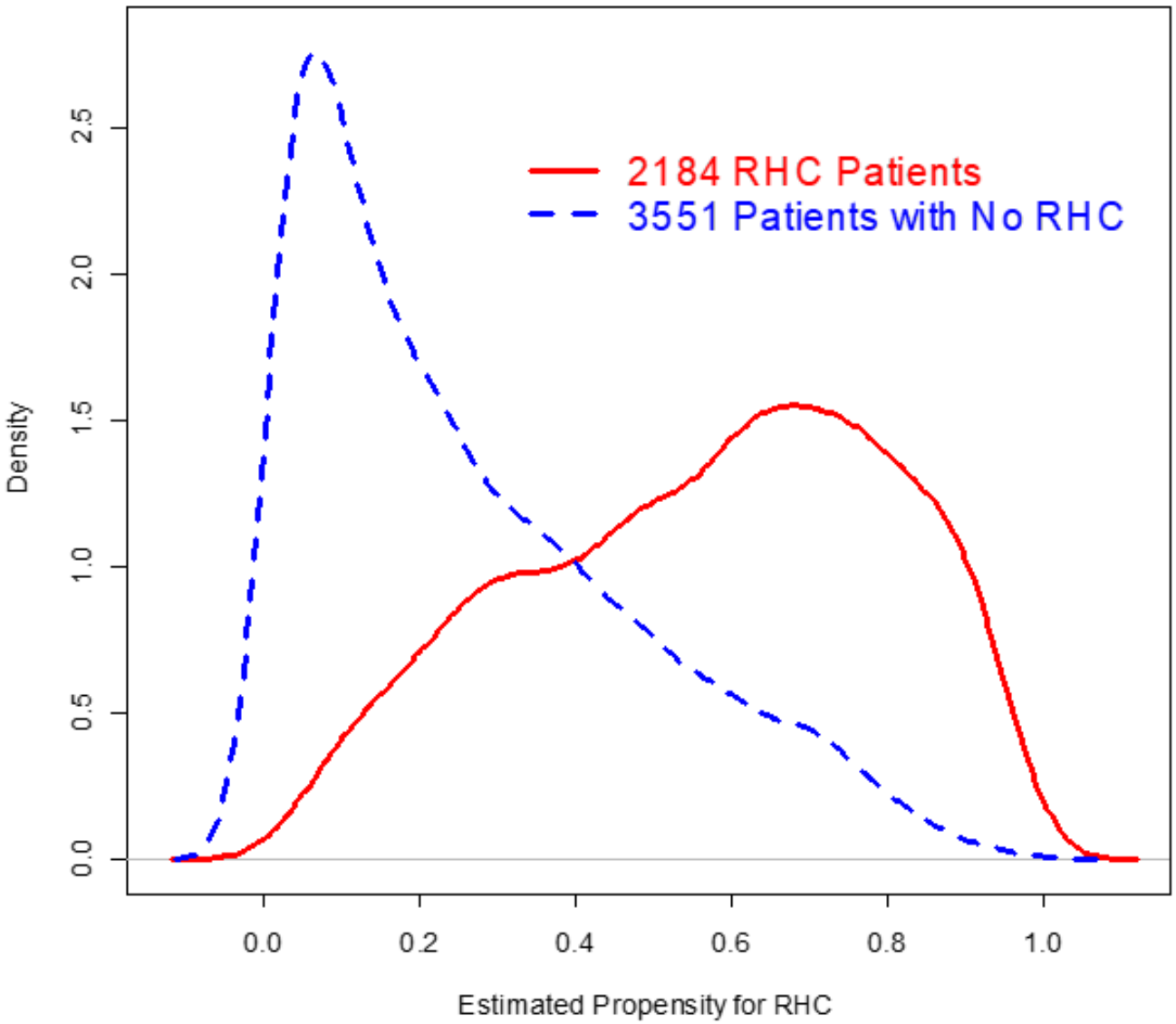
RHC patients were less likely to

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

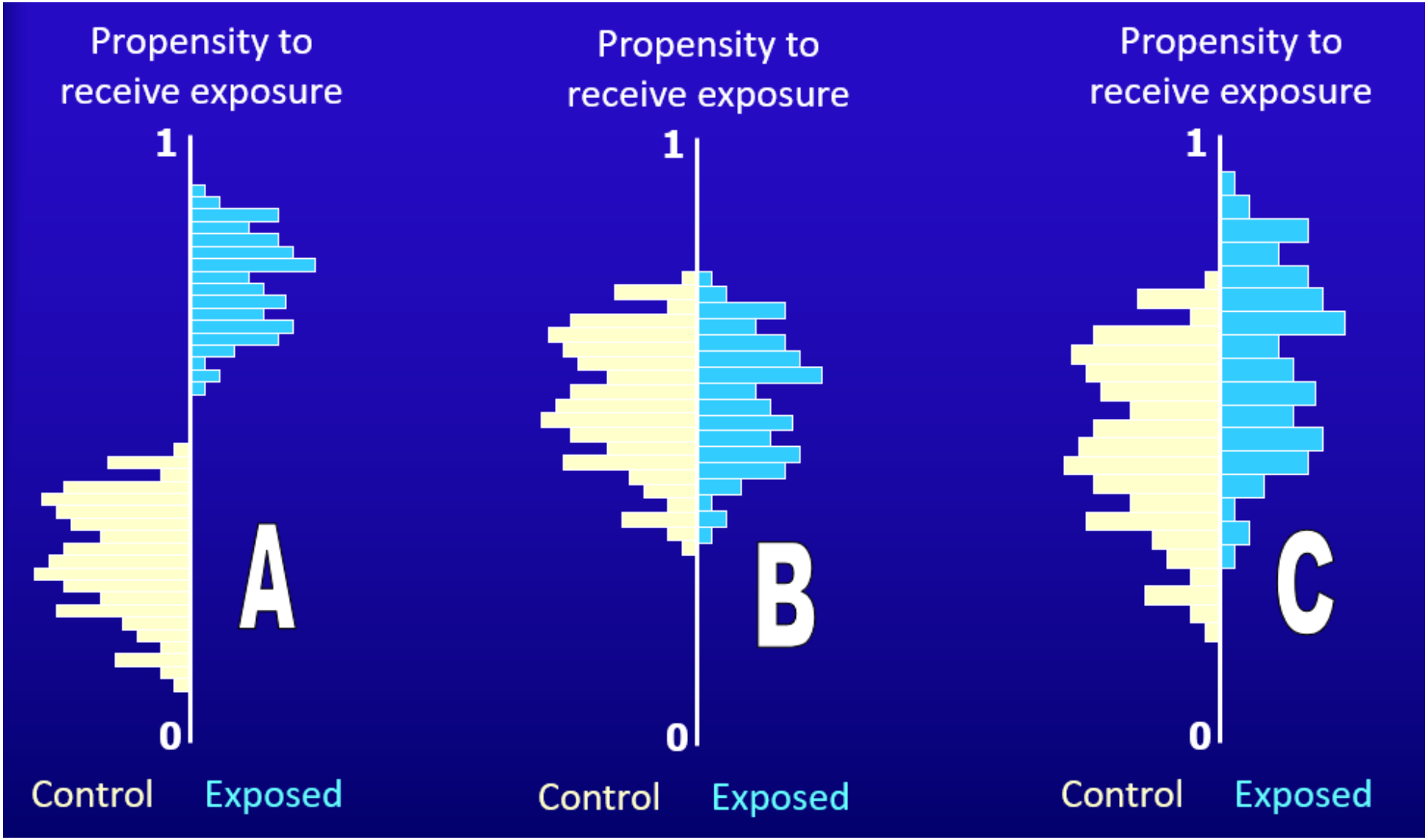
RHC patients had significantly

- 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

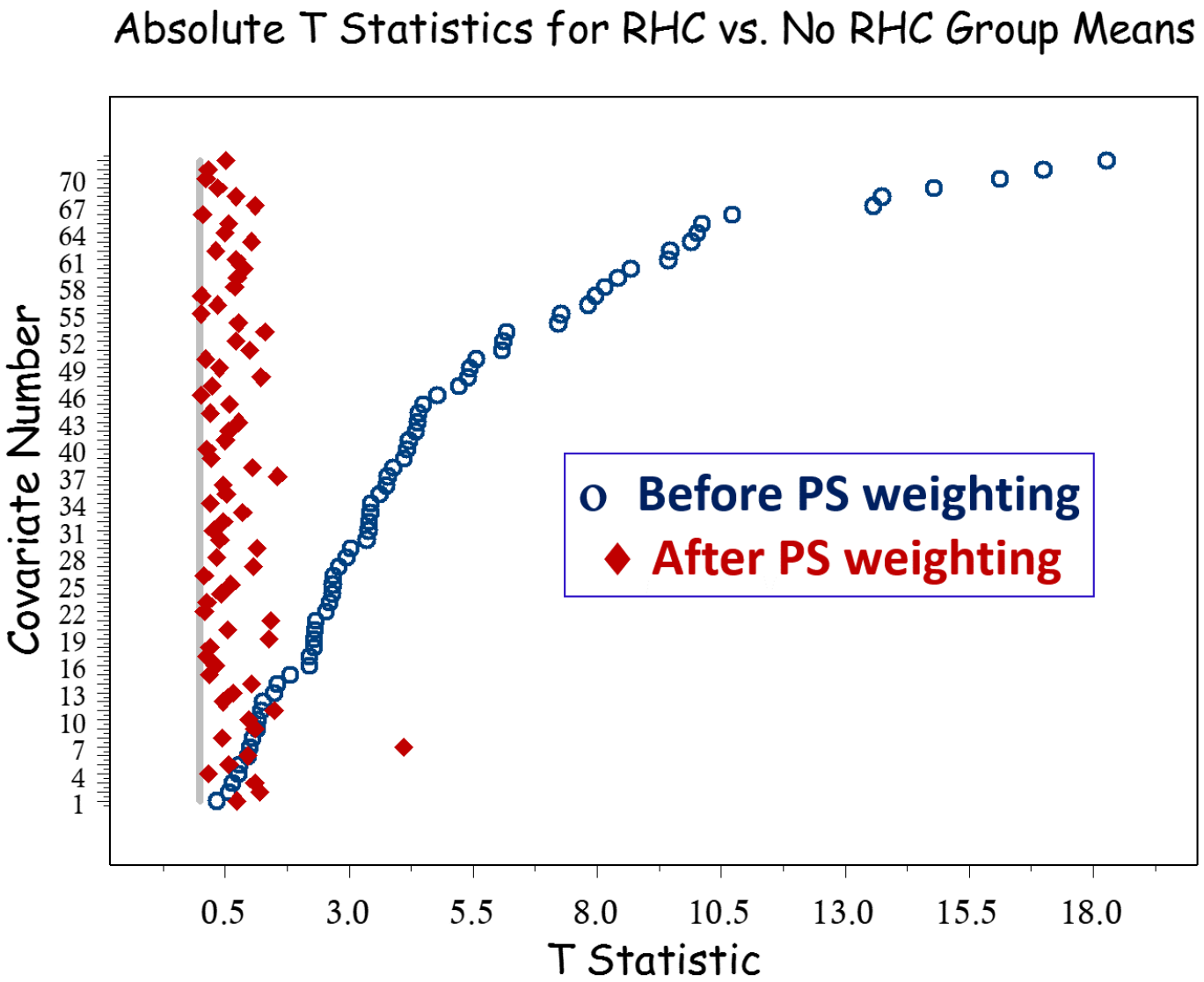
- 5,735 hospitalized patients in SUPPORT study
  - 2,184 treated (RHC) and 3,551 controls (no RHC).

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

- PS model estimated by Hirano and Imbens[10] 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.

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

Results of this Weighting Approach



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.

What Propensity Scores Can and Cannot Do

- If we match exposed subjects to controls with similar propensity scores, we can behave as if they had been randomly assigned to exposures.
- Or, if we use weighting with or without additional regression to adjust for propensity to get treatment, we can compare exposed subjects 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 we'll never know it.

Actually Doing Propensity Score Analyses in R

Slides to Pull from `useful_ppt.pptx`

Books: 130, 131, Leite's book, Paul's three books

Articles: 132 and add a new one or two (King + the new thing on Love plots), OS Journal, HSR&OM

Links: 133 Elizabeth Stuart, Rand TWANG, Peter Austin, other speakers in this series

Packages 134, 135

Sensitivity Analysis 148-170

182 - variable selection in PS model

- King et al. (2017) argues somewhat persuasively that the use of propensity score matching to emulate an RCT isn't a good idea, even if using other propensity-score adjustments is a good idea.

Closing Thoughts

A Few Advantages of Propensity Scores

- Results can be persuasive even to audiences with limited statistical training.
- Though estimating the PS requires some care, the comparability of exposed and control patients can be verified simply.
- PS methods address selection bias well.
- PS methods may be combined with other sorts of adjustments.

Strategic Issues

- How can we make our investigations compelling to our intended audience?
- Why is this hard?
  - Audience is not focused on statistical techniques
  - Audience may have limited training in statistics
- Why is this important?
  - Who makes key policy decisions?
  - Who needs to be convinced by the evidence?

Strategic Issues in Observational Studies

- Design observational studies
  - Exert as much experimental control as possible, carefully consider the selection process, and anticipate hidden biases
- Focus on simple comparisons
  - Increase impact of results on consumers
- Compare subjects who looked comparable prior to treatment
- Use sensitivity analyses to delimit discussions of hidden biases due to unobserved covariates

See Rosenbaum, 2002, 2010 and 2017

Some Cautions and Limitations

- Hidden Bias: Beware unmeasured covariates which affect outcomes and/or assignment.
  - Sensitivity Analysis helps quantify the problem
- This is a reasonable method with fairly large samples.
- Matching vs. stratification vs. adjustment methods
- Options narrow as an investigation proceeds.
- Sadly, though OS work cries out for design, we're often working with secondary data, where we have fewer options

What Should Always be done in an OS, and often isn't?

1. Collect data so as to be able to model selection
2. Demonstrate need for adjustment - selection bias
3. Carefully record intervention time - adjust only for things present before or at time of intervention.
4. Ensure baseline characteristic overlap [comparability]
5. Check baseline characteristic balance after adjustment
6. Specify relevant post-adjustment population with care
7. Estimate intervention effect in light of adjustment
8. Estimate sensitivity of results to potential hidden bias