

Propensity Scores: What Do They Do, How Should I Use Them, and Why Should I Care?

Thomas E. Love, PhD

Center for Health Care Research & Policy

Case Western Reserve University

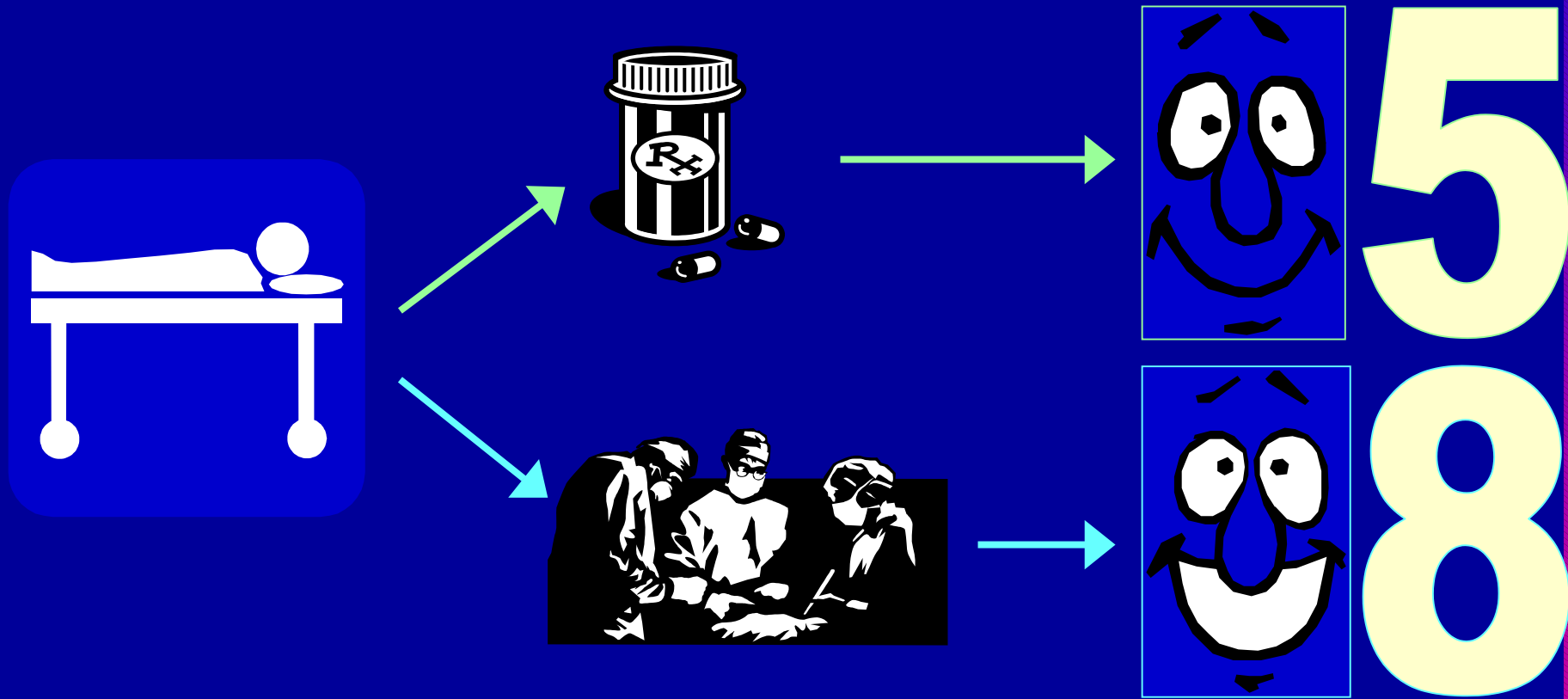
thomaslove@case.edu

ASA Cleveland Chapter, December 2003

The Plan for This Evening

- Why should I care about propensity scores?
 - They provide a means for adjusting for selection bias in observational studies of causal effects.
- What do propensity scores do?
 - They summarize all of the background (covariate) information about treatment selection into a scalar.
- How should I use propensity scores?
 - Whenever you want to make a causal inference in comparing exposures, especially when you have a lot of information on how the exposures were selected.

Looking for Causal Treatment Effects



Surgery Drug Treatment Effect =

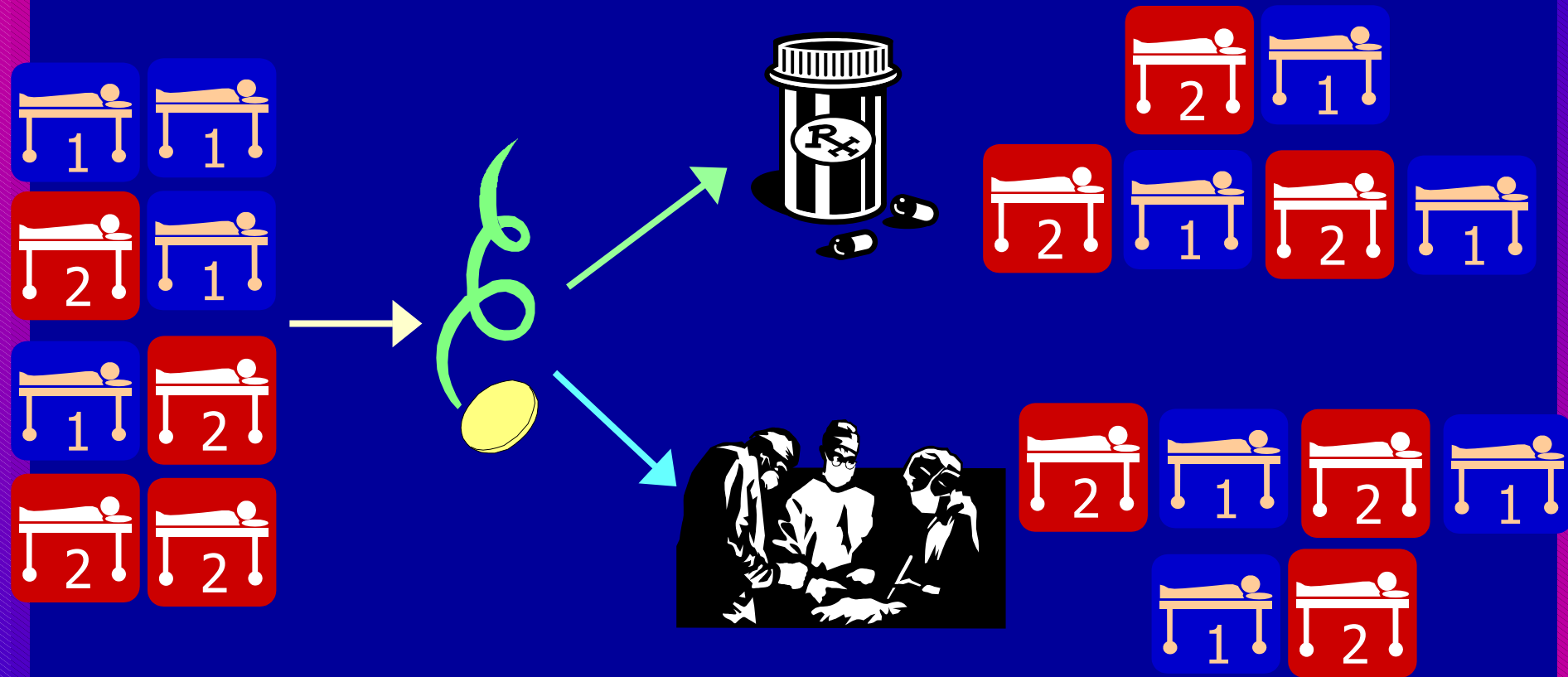


Ceteris Paribus

Other Things Being Equal

- Suppose we are comparing a treated group to a control group, and we want to know if the treatment has a causal effect on the outcome.
- To be fair, we must compare treated and controls who are similar in terms of everything that affects the outcome, except for the receipt of treatment.
- How do we, as statisticians, typically recommend that people do this?

Importance of Randomization



Randomization ensures that subjects receiving different treatments are comparable.

RCT Subject Selection

MRFIT Excluded 96.4% of potential eligibles

361,662 men
Age: 35 - 57

25,545

12,866

RANDOMIZATION

6428

6438

EXCLUSIONS

- Low Risk of CHD
- Hx of MI
- Diabetes
- Geographic Mobility
- Cholesterol > 350
- DBP > 115

336,117

EXCLUSIONS

- Body Wgt > 150%
- Angina
- Evidence of MI
- Strange Diet

12,679

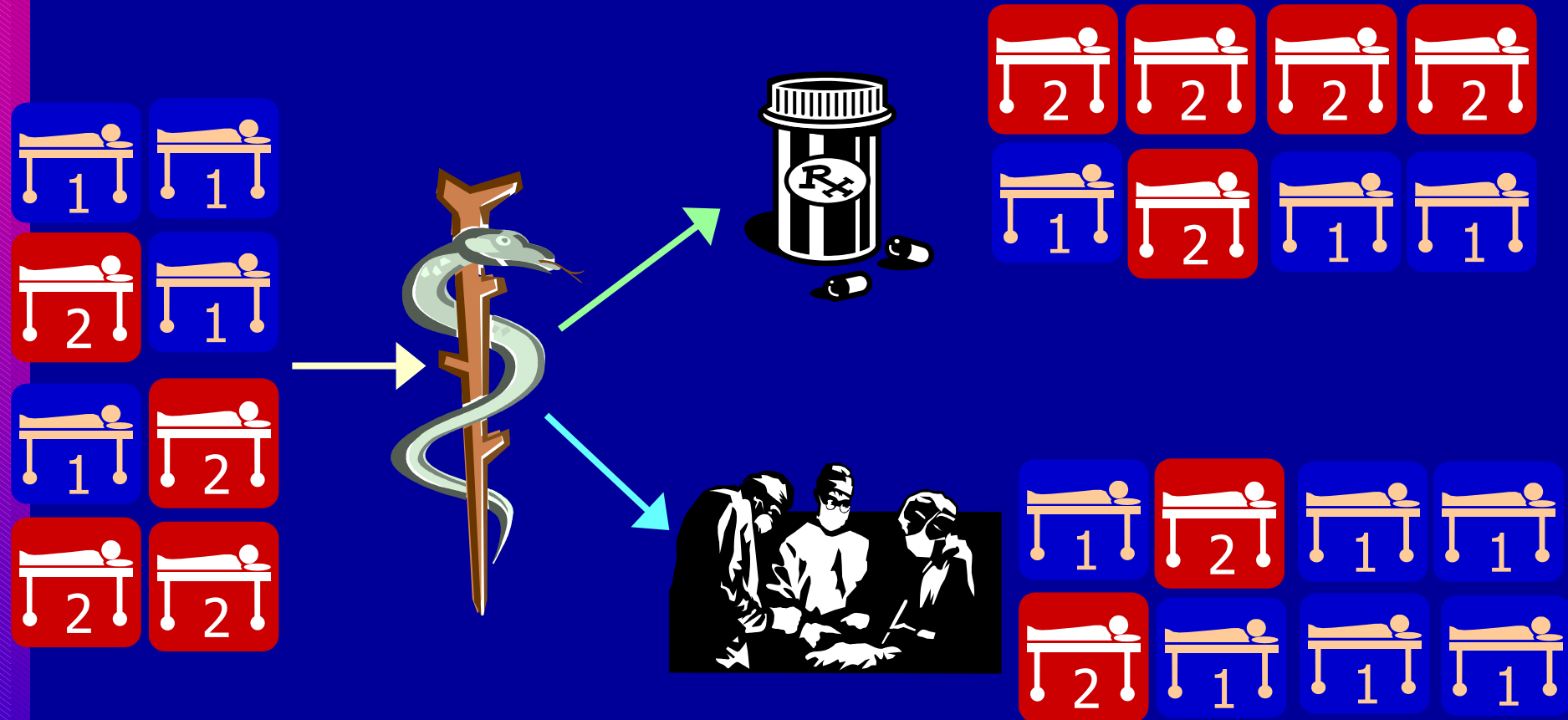
How Are Experiments Designed?

- Most important feature of experiments:
 - We must decide on the way data will be collected before observing the outcome.
- Experiments are “honest”
 - Want to use statistics and data to illuminate our understanding, not just to support our priors.
 - If we could try out lots of designs and look at what that does for the answer, we would, and we’d then “always” confirm our prior beliefs.

We Usually Assign Treatments At Random

- Except sometimes, we can't.
- Or we just don't.
- And sometimes, these are the situations we care most about.
- So how can we assess causal effects in observational studies?

Observational Studies



In an observational study, the researcher **does not** randomly allocate the treatments.

On Designing Observational Studies

- Exert as much experimental control as possible
 - Carefully consider the selection process
 - Actively collect data to reveal potential biases
- “Care in design and implementation will be rewarded with useful and clear study conclusions ... Elaborate analytical methods will not salvage poor design or implementation of a study.” -- NAS report (Rosenbaum (2002) p. 368)

"Hormone Replacement Therapy Should Be Recommended for Nearly All Women"

Col NF, Eckman MH, *et al.* (1997) *JAMA*, 277: 1140-47.

CHD risk for patients using HRT was 0.60 times as high as the risk for patients not using HRT.

"Do not use estrogen/progestin to prevent chronic disease"

Fletcher SW, Colditz G. (2002) *JAMA*, 288: 366-67.

Manson JE *et al.* for WHI Investigators (2003) *NEJM*, 349: 523-534.

CHD risk for patients using HRT was 1.29 times as high as the risk for patients not using HRT.

What Propensity Scores Are, and Why They Are Important

- We can gather up all the background info that we have on our subjects before treatments are assigned, that might plausibly affect which treatment they get
- We build a model to predict the probability that they will receive the treatment instead of the control.

$$\text{Propensity Score} = \Pr(\text{Treated} \mid \text{background info})$$

- Groups of subjects with similar propensity scores can then be expected to have similar values of all of the background information, in the aggregate.

The Propensity Score

Pr(treatment given covariates)

- **Definition:** The conditional probability of receiving a given exposure (treatment) given a vector of measured covariates.
- Usually estimated using logistic regression:

$$\ln\left(\frac{PS}{1-PS}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

$$PS = \frac{\exp(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}$$

I
II
III
IV
V

VI
VII
VIII
IX
X



The Ten Commandments of Propensity Model Development



Thou Shalt Value Parsimony

Thou Shalt Examine Thy
Predictors For Collinearity

Thou Shalt Test All Thy Predictors
For Statistical Significance

Thou Shalt Have Ten Times
As Many Subjects As Predictors

Thou Shalt Carefully Examine Thy
Regression Coefficients (Beta Weights)



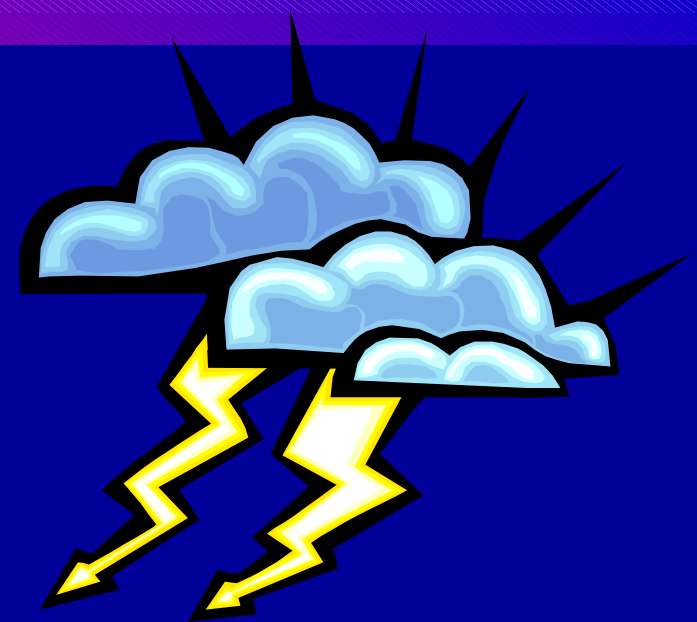
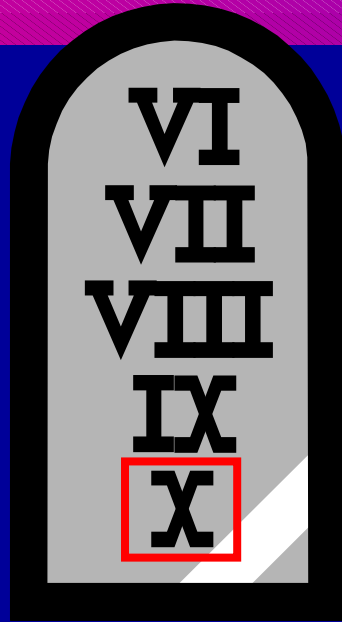
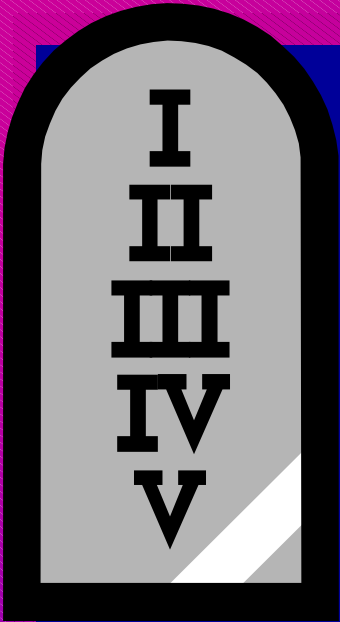


Thou Shalt Perform Bootstrap
Analyses To Assess Shrinkage

Thou Shalt Perform Regression
Diagnostics and Examine Residuals
With Care

Thou Shalt Hold Out A Sample of
Thy Data for Cross-Validation

Thou Shalt Perform External
Validation on a New Sample of Data



Thou Shalt Ignore
Commandments 1 through 9...
And Instead Simply **Ensure**
That The Model Adequately
Balances The Covariates

Should Adjustments Be Made for All Observed Covariates?

- If not, how should covariates be selected?
- No real reason to avoid adjustment for a true covariate – a variable describing subjects before treatment.
- In practice, though, this can increase both cost and complexity unnecessarily.
- There are issues of data quality and completeness to consider.

Screening for Covariates in the Propensity Score Model

- Most common method **ERRIBLE IDEA!**
 - Compare treated to non-treated on many covariates.
 - Adjust only for significant differences.
- No reason that absence of significance implies imbalance is small enough to be ignored.
- Doesn't consider covariate-to-outcome relationship.
- This process considers covariates one at a time, while the PS adjustments will control the covariates simultaneously.

OK, What Should We Do?

- Give the data multiple opportunities to call attention to potential problems.
- Select a tentative list of covariates for adjustments using problem knowledge and exploratory comparisons of treatment groups.
- Select tentative adjustment method and apply it to the covariates excluded from the list, identifying large imbalances after adjustment.
- Reconsider the tentative list in light of this.

What Propensity Scores Can and Cannot Do, in a Slide

- If we match 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 any baseline differences on selection to treatment or control.
- **But** if our propensity model misses an important reason why subjects are selected to treatment or control, we'll be in trouble.

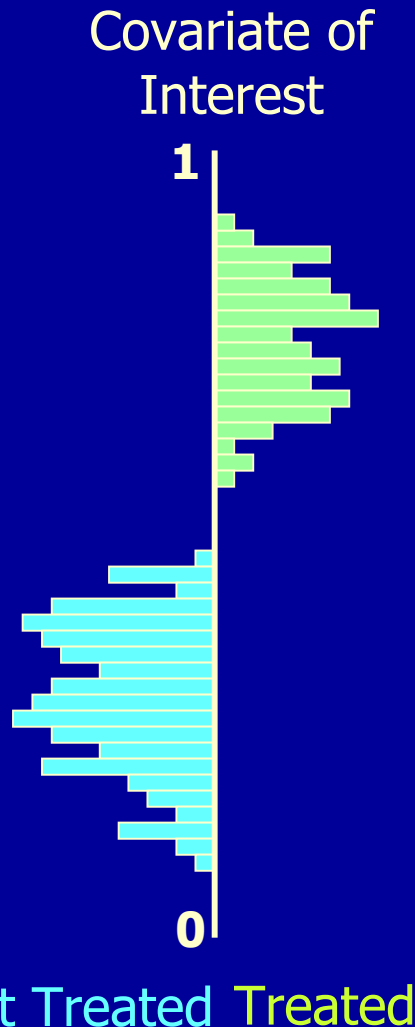
Rich and Poor Covariate Sets

- With a rich set of covariates, adjustments for hidden covariates may be less critical.
- With less rich covariate sets, we may need to do more – say, try to find an instrument.
- Conclusion after the initial design stage may be that the treatment and control groups are too far apart to produce reliable effect estimates without heroic modeling assumptions.

Why Work This Hard in the Initial Stage of Design?

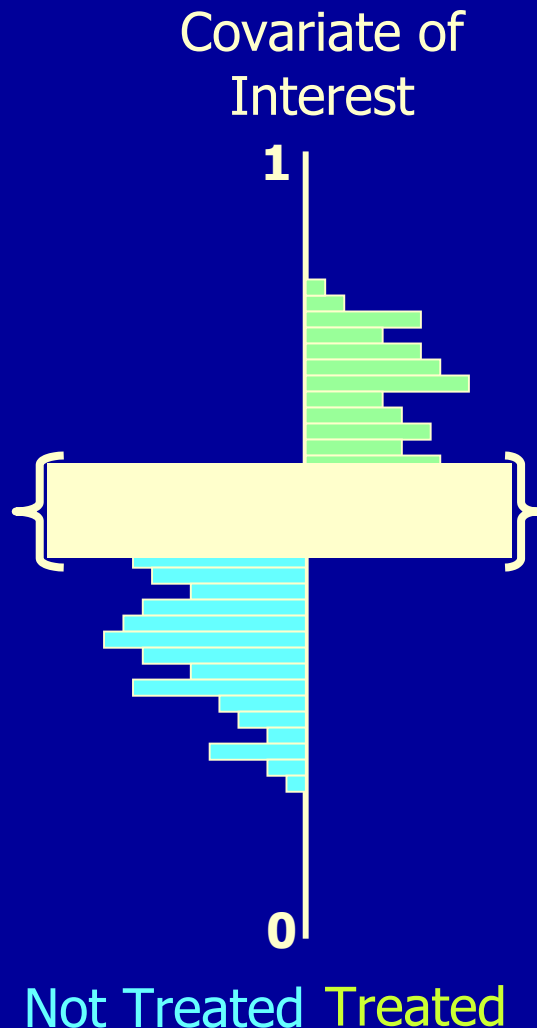
- **No harm, no foul.** Since no outcome data are available to the PS, nothing based on the PS here biases estimation of treatment effects.
- Balancing covariates / PS makes subsequent model-based adjustments **more reliable**.
 - Model adjustments can be extremely unreliable when the treatment groups are far apart on covariates.
 - Helps covariance, relative risk, IV adjustments, etc.

How Much Overlap In The Covariates Do We Want?



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

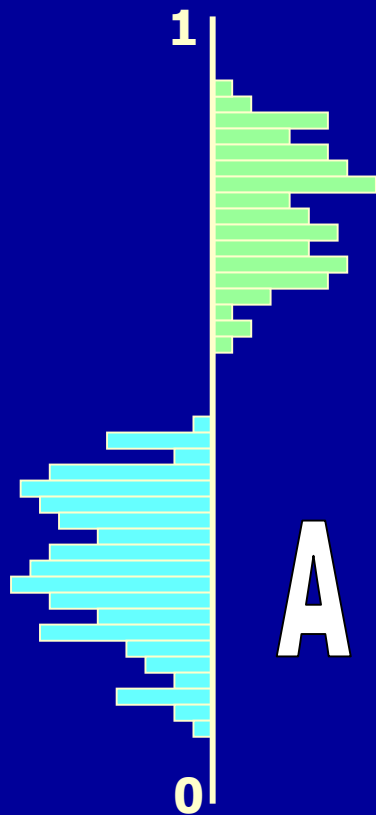
What if Treated and Untreated Groups Overlap, but minimally?



- Not much help.
- The information available to infer treatment effect will reside almost entirely in the few patients who overlap.
- Need to think hard about whether useful inferences will be possible.

How Much Overlap In The Propensity Scores Do We Want?

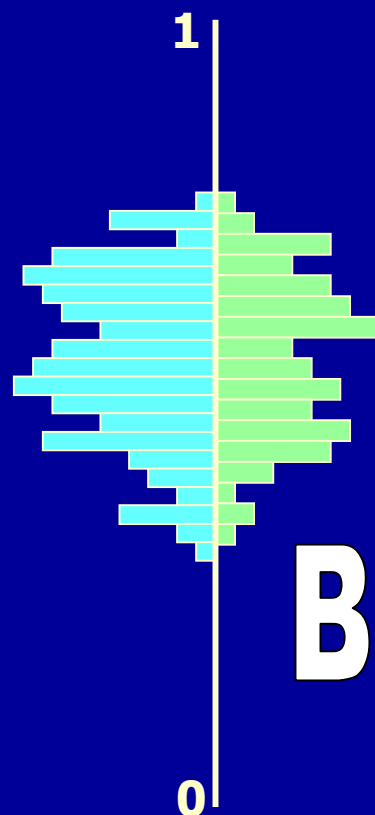
Propensity to receive treatment



A

Not Treated Treated

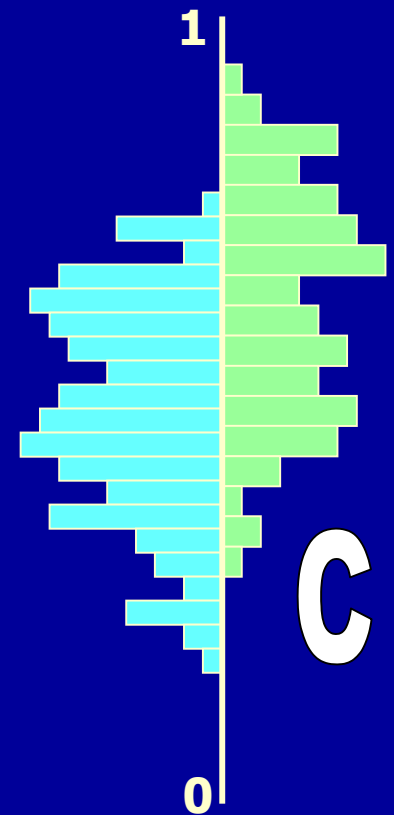
Propensity to receive treatment



B

Not Treated Treated

Propensity to receive treatment



C

Not Treated Treated

Aspirin Use and Mortality

- 6174 consecutive adults undergoing stress echocardiography for evaluation of known or suspected coronary disease.
- 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)

Baseline Characteristics By Aspirin Use (in %) (before matching)

| Variable | Aspirin (n = 2310) | No Aspirin (n = 3864) | P value |
|-------------------------------|-----------------------|--------------------------|---------|
| Men | 77.0 | 56.1 | < .001 |
| Clinical history: diabetes | 16.8 | 11.2 | < .001 |
| hypertension | 53.0 | 40.6 | < .001 |
| prior coronary artery disease | 69.7 | 20.1 | < .001 |
| congestive heart failure | 5.5 | 4.6 | .12 |
| Medication use: Beta-blocker | 35.1 | 14.2 | < .001 |
| ACE inhibitor | 13.0 | 11.4 | < .001 |

- Baseline characteristics appear very dissimilar: 25 of 31 covariates have $p < .001$, 28 of 31 have $p < .05$.
- Aspirin user covariates indicate higher mortality risk.

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 $p = .01$]

Using Standardized Differences to Measure Covariate Balance

- Standardized Differences greater than 10% in absolute value indicate serious imbalance

$$d = \frac{100(\bar{x}_{Treatment} - \bar{x}_{Control})}{\sqrt{\frac{s_{Treatment}^2 + s_{Control}^2}{2}}} \quad \text{for continuous variables}$$

$$d = \frac{100(p_{Treatment} - p_{Control})}{\sqrt{\frac{p_T(1-p_T) + p_C(1-p_C)}{2}}} \quad \text{for binary variables}$$

|Standardized Differences| > 10% Indicate Serious Imbalance

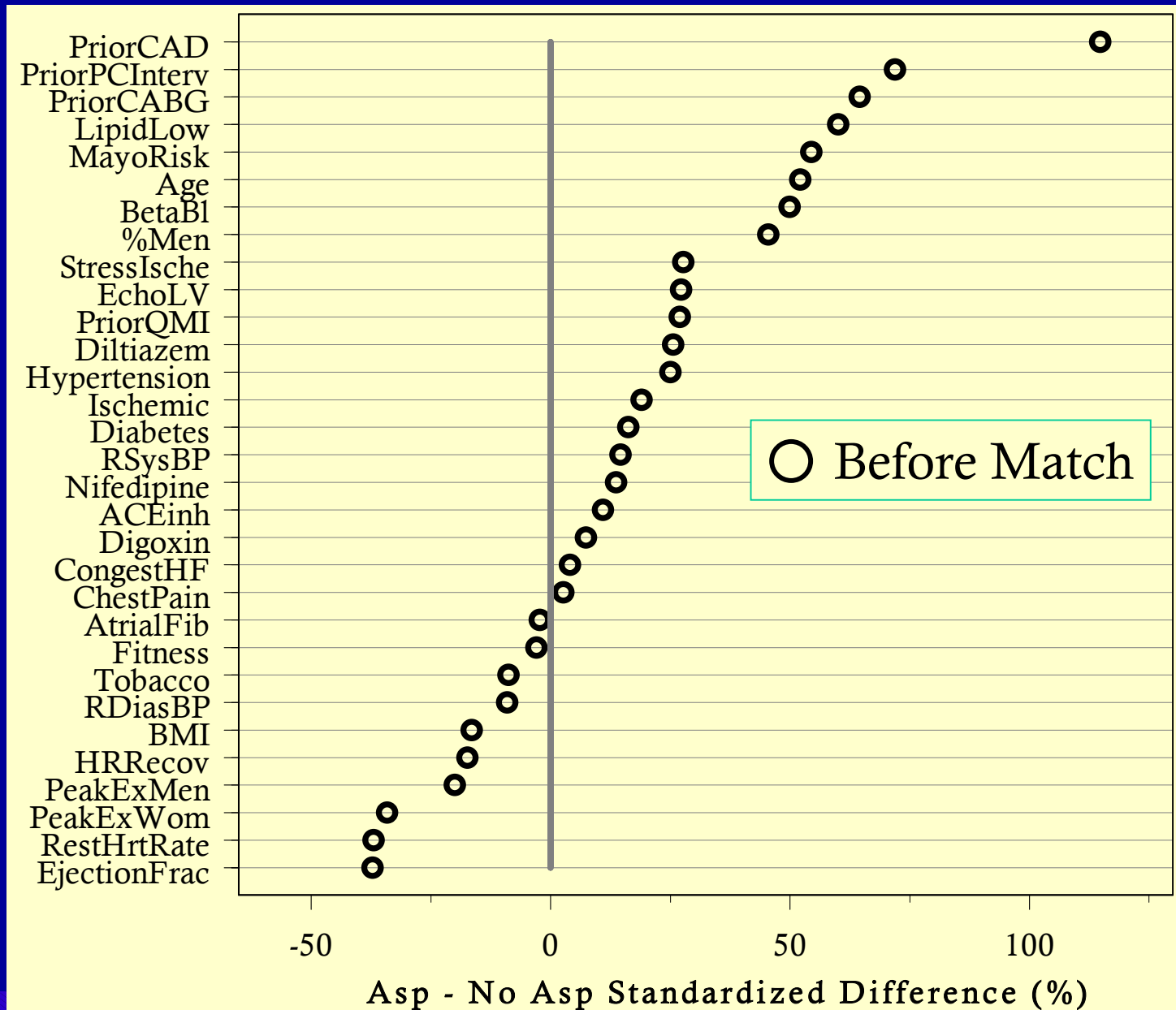
Before Match:

- 811/2310 (35.1%) Aspirin users used β -blockers
- 550/3864 (14.2%) non-Aspirin users used β -blockers
- Standardized Difference is 49.9%
- P value for difference is < .001

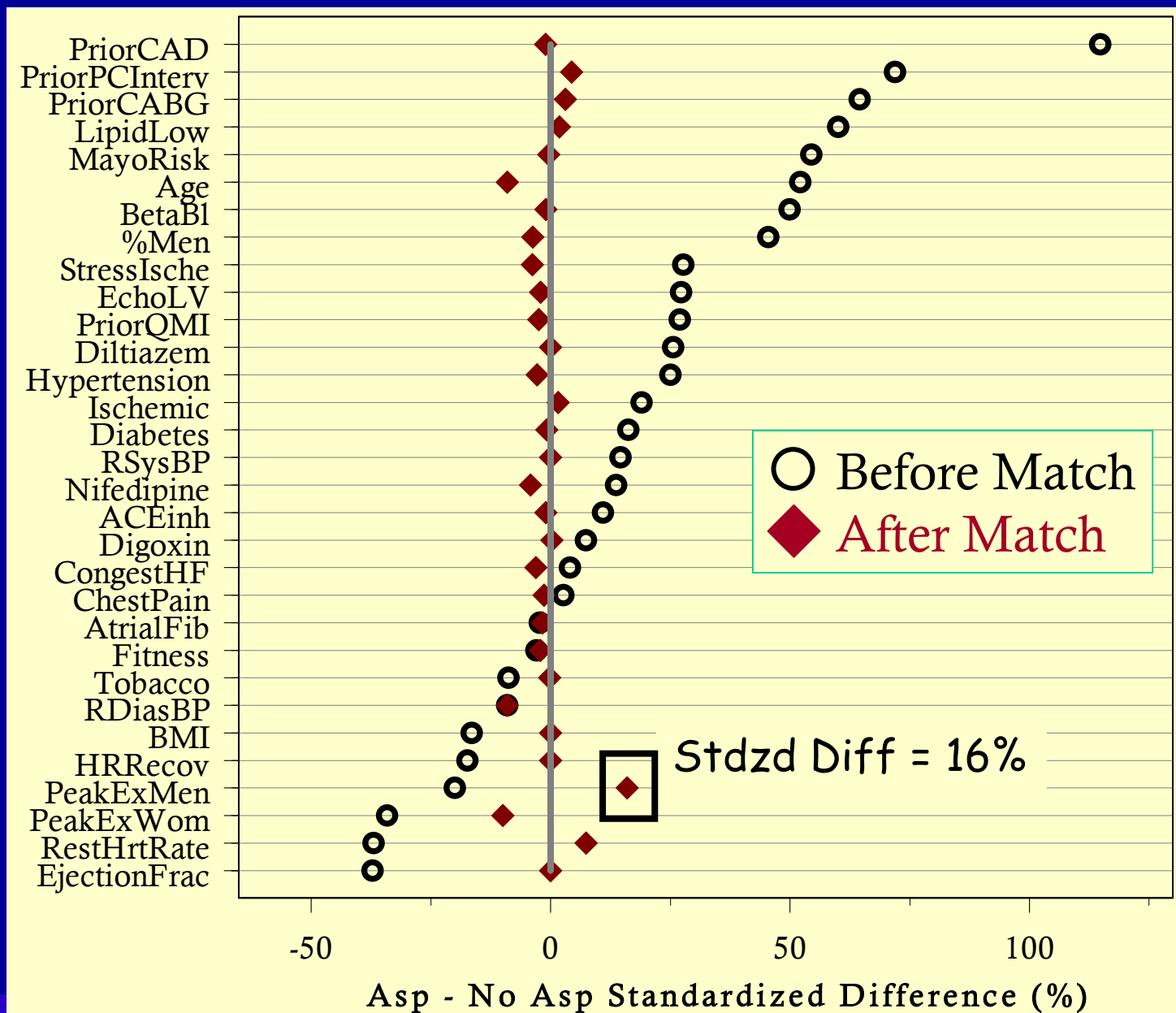
After Match:

- 352/1351 (26.1%) Aspirin users used β -blockers
- 358/1351 (26.5%) non-Aspirin users used β -blockers
- Standardized Difference is –1.0%
- P value for difference is .79

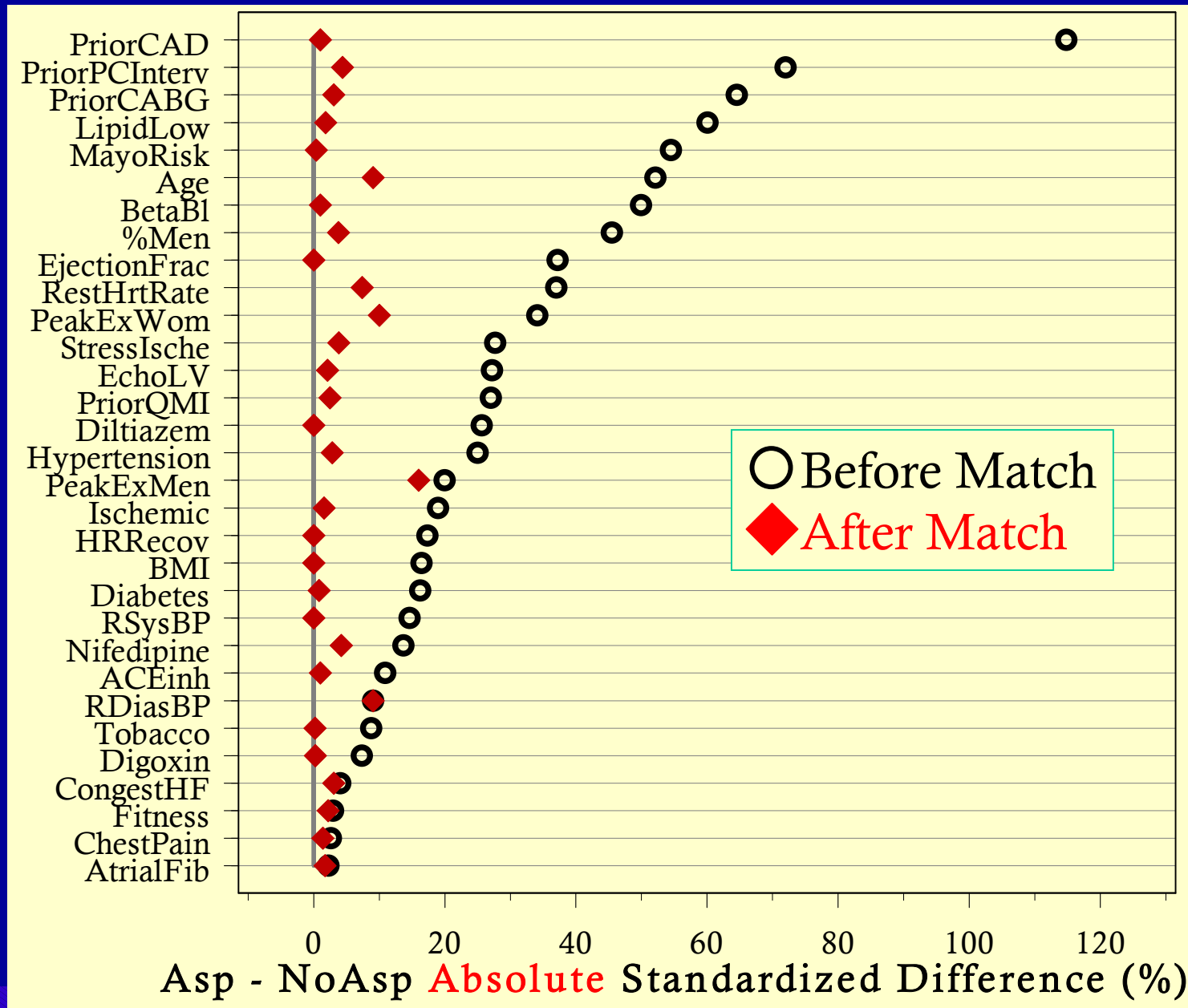
Covariate Balance for Aspirin Study



Covariate Balance for Aspirin Study



Absolute Standardized Differences

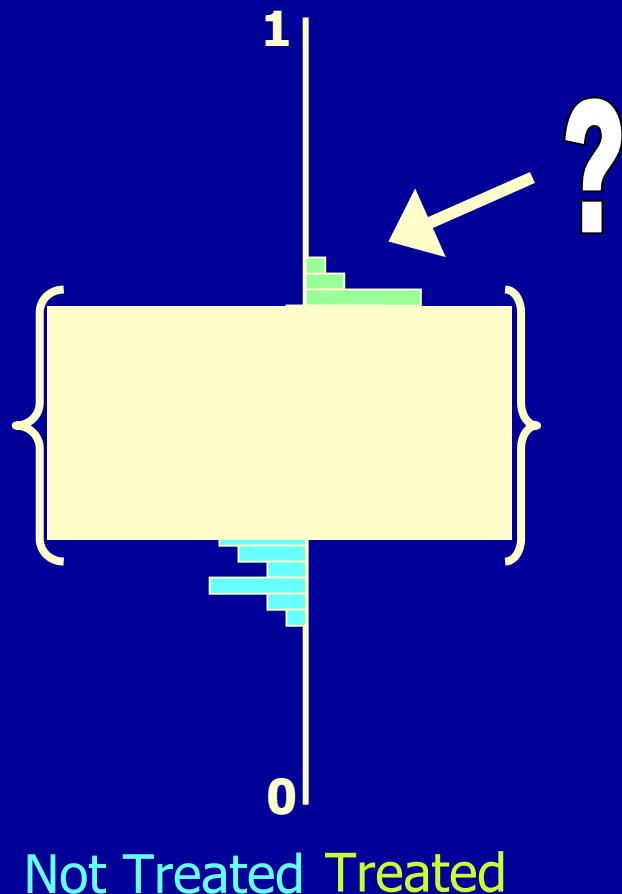


Incomplete vs. Inexact Matching

- Trade-off between
 - Failing to match all treated subjects (**incomplete**)
 - Matching dissimilar subjects (**inexact** matching)
- There can be a severe bias due to incomplete matching
 - it's often better to **match all treated subjects**, then follow with analytical adjustments for residual imbalances in the covariates.
- In practice, concern has been inexactness.
- Certainly worthwhile to define the comparison group and carefully explore why subjects match.

What if Treated and Untreated Groups Don't Overlap Completely?

Propensity Score



- Inferences for the causal effects of treatment on the subjects with no overlap cannot be drawn without heroic modeling assumptions.
- Usually, we'd exclude these treated subjects, and explain separately.

Which Aspirin Users Get Matched?

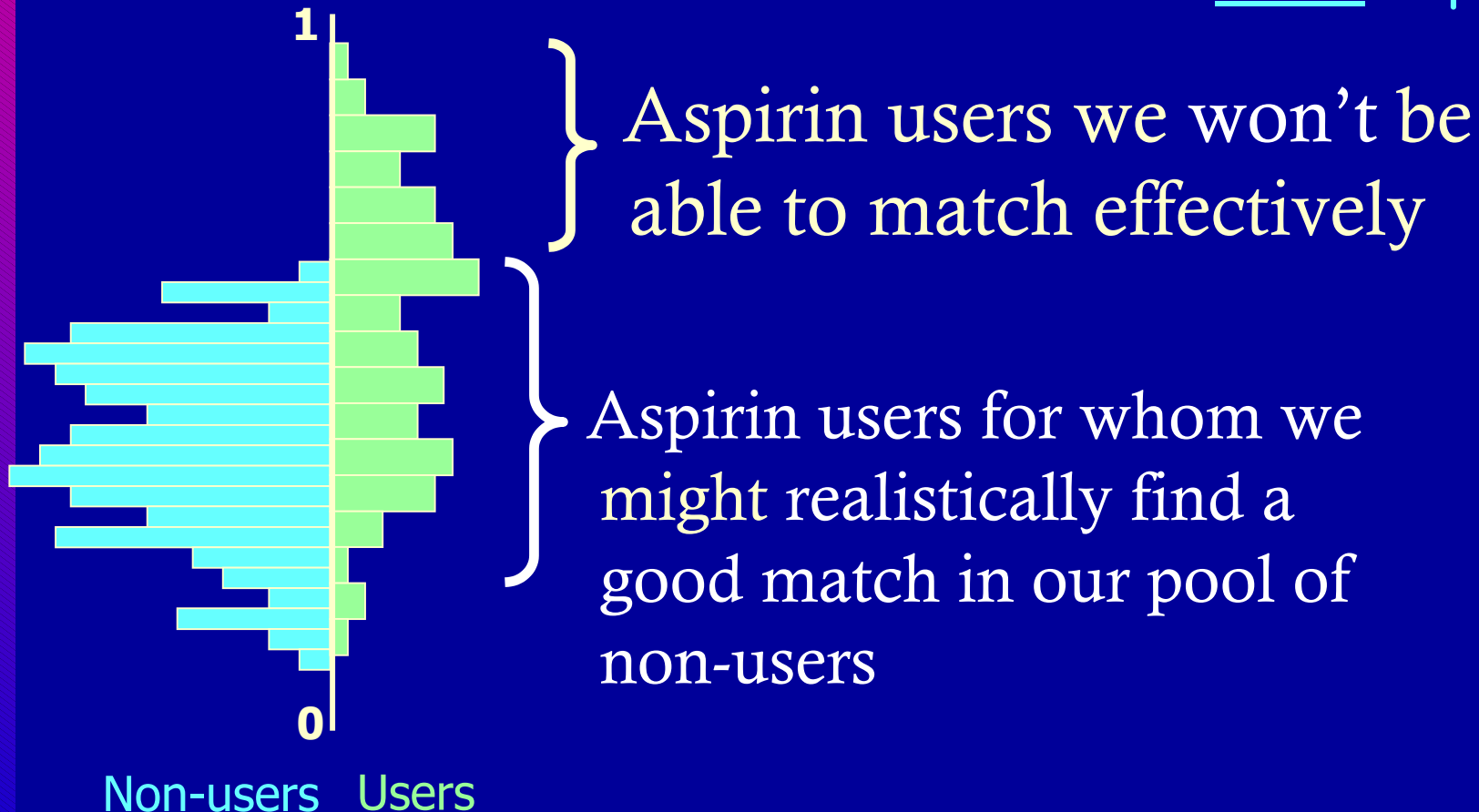
- Generally, characteristics of unmatched aspirin users tend to indicate high propensity scores.
 - Overall, 37% of patients were taking aspirin.
 - The rate was much higher in some populations...
67% of Prior CAD patients were taking aspirin.
 - So prior CAD patients had higher propensity scores for aspirin use.
 - Of the unmatched aspirin users, 99.8% (957/959) had prior coronary artery disease.
 - So it's likely that the unmatched users tended towards larger propensity scores than the matched users.

Who's Getting Matched Here?

Where Do The Propensity Scores Overlap?

Propensity to
Use Aspirin

Caveat: This simulation depicts
what often happens.



Matching with Propensity Scores

- 1351 aspirin subjects (58%) matched to non-aspirin subjects – big improvement in covariate balance. Matched group looks like an RCT...
- Matching still incomplete, but results on PS matched group mirrored the results for the covariate-adjusted group as a whole...
- Resulting matched pairs analyzed using standard statistical methods, e.g. Kaplan-Meier, Cox proportional hazards models.

Estimating The Hazard Ratios

| Approach | n | Hazard Ratio | 95% CI |
|---|------|--------------|-------------|
| Full sample, no adjustment | 6174 | 1.08 | (.85, 1.39) |
| Full sample with no PS, adjusted for all covariates | 6174 | 0.67 | (.51, .87) |
| PS-Matched sample | 2702 | 0.53 | (.38, .74) |
| PS-Matched, adjusted for PS and all covariates | 2702 | 0.56 | (.40, .78) |

- During follow-up 153 (6%) of the 2702 propensity score-matched patients died.
- Aspirin use was associated with a lower risk of death in matched group (4% vs. 8%, $p = .002$).

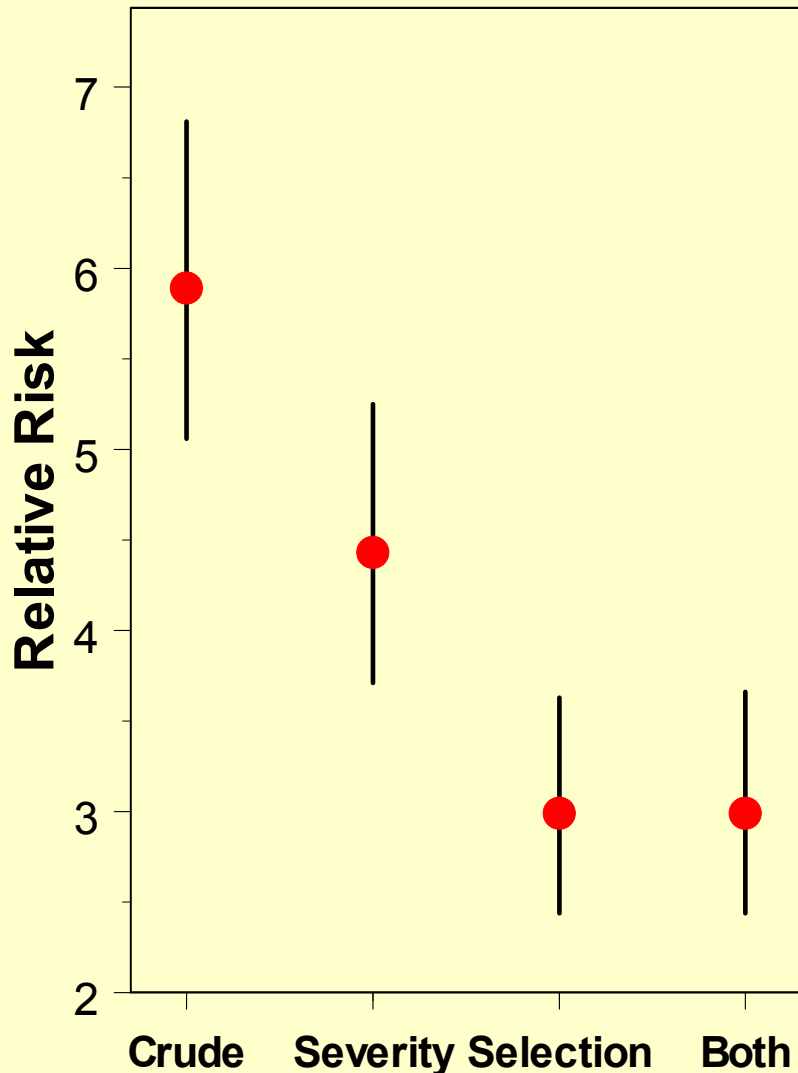
Aspirin Conclusions / Caveats

- Patients included in this study may be a more representative sample of “real world” patients than an RCT would provide.
- PS matching is still not randomization: can only account for the factors measured, and only as well as the instruments can measure them.
- No information on aspirin dose, aspirin allergy, or duration of treatment, or on medication adjustments.

Propensity Score Matching in the Design of an Observational Study

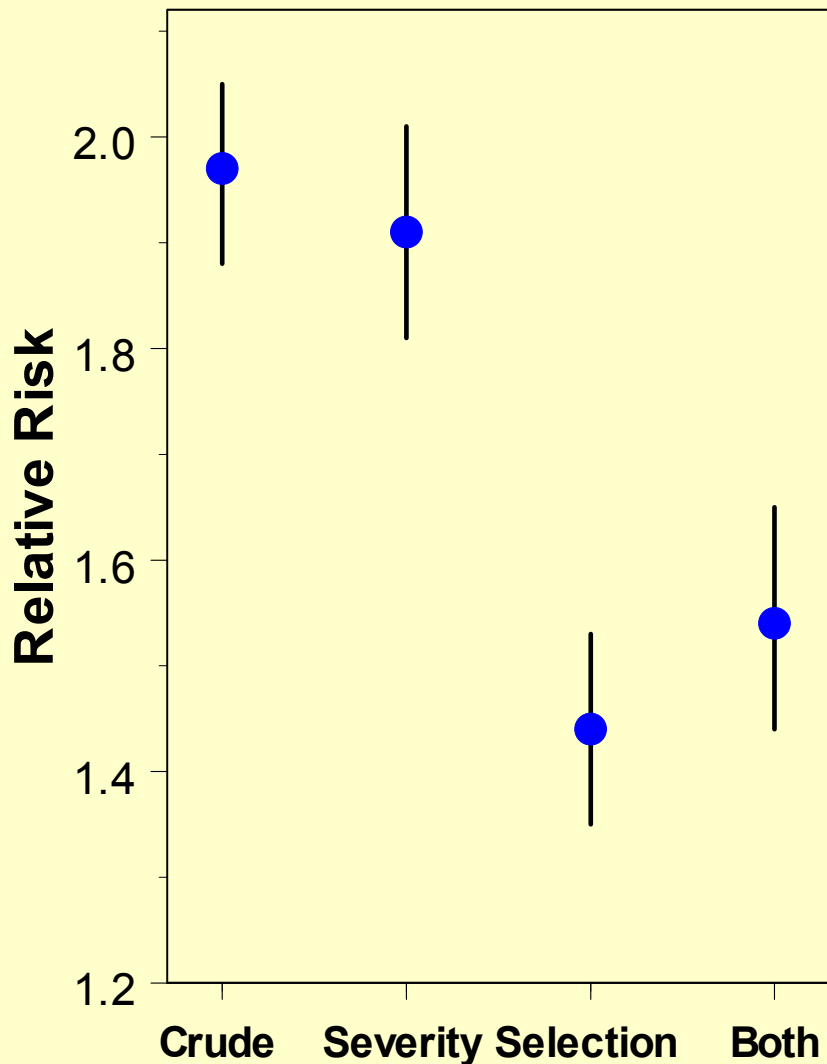
- Pair up treated and control subjects with similar values of the propensity score, discarding all unmatched units.
- Not limited to 1-1 matches – can do 1-many.
- Can find an optimal match, without discarding any units, then follow with adjustments.
 - Technically more valid, but difficult sell in practice.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

Adjusting for Severity vs. Adjusting for Selection vs. Both



- Effect of pneumonia as a complication of stroke (on 30-day mortality)
 - Katzan, Cebul et al. (2003) *Neurology*.
- Relative risk of death by 30 days for patients with pneumonia vs. patients without.
- Severity – risk-adjusted
- Selection – PS-adjusted

Adjusting for Severity vs. Adjusting for Selection vs. Both



- Effect of rehabilitation on discharge to home from nursing home
 - Murray et al. (2003)
Arch Phys Med Rehab.
- Severity – risk-adjusted
- Selection – PS-adjusted
- Selection bias blunts the rehab effect size.

On Planning an Observational Study (Rosenbaum, 2002)

- A convincing OS is the result of active observation, a search for those rare circumstances in which tangible evidence may be obtained to distinguish treatment effects from the most plausible biases.
- Experimental control is replaced in a good OS by careful choice of environment. Design is crucial!
 - Options narrow as an investigation proceeds.
 - These are reasonable methods with large samples, especially if we have a good selection model using multiple covariates.

What should always be done in an OS ... and often isn't?

- Collect data so as to be able to model selection
- Demonstrate selection bias – need for PS
- Ensure covariate overlap for comparability
- Evaluate covariate balance after PS application
- Specify relevant post-adjustment population carefully
- Model or estimate treatment effect in light of PS adjustment / matching / stratification
- Estimate sensitivity of results to potential hidden bias

What I Discussed This Evening According to the Abstract

1. Investigating the issue of selection bias well, and how selection bias can affect results,
2. Assessing the degree of “overlap” with regard to background characteristics in the two exposure groups, to determine if an effect can be sensibly estimated using the available data,
3. Estimating treatment effects adjusting for observed differences in background characteristics using propensity scores,
4. Assessing whether the propensity score worked well, in the sense of producing “fair” comparison groups.

Where You Can Go for More Information

- <http://www.chrp.org/propensity>
- Free issue of Health Services and Outcomes Research Methodology (December 2001 v2 issues 3-4) – go to <http://www.kluweronline.com/issn/1387-3741>
 - I especially recommend the article by DB Rubin (using PS in designing a complex observational study) and the article by Landrum and Ayanian (PS vs. Instrumental Variables)
- Rosenbaum PR (2002) *Observational Studies*, Springer.
- Email me at thomaslove@case.edu