# 500 Class 06

https://thomaselove.github.io/500-2023/

2023-02-23

# What's in these Slides?

- Matching Approaches and Austin 2014 (also in Slides Set 5)
- Feinstein's Model for Research Architecture (also in Slides Set 5)
- Designing Observational Studies (Rubin, 2001)
- Discussion of Rosenbaum Chapters 5 and 6
- Discussion of Project Proposals

# Section 1

## Matching Approaches (discussion built on Austin, 2014)

# 1:1 Greedy Matching

Greedy (or nearest neighbor) matching selects a treated subject and then selects as a matched control subject the untreated subject whose propensity score is closest to that of the treated subject. If multiple untreated subjects are equally close to the treated subject, one of these untreated subjects is selected at random, typically. Options include:

1. Select treated subjects from highest to lowest propensity score.
2. Select treated subjects from lowest to highest propensity score.
3. Select sequentially treated subjects in the order of the best possible match.
    - First treated subject is the one who is closest to an untreated subject.
    - Second treated subject is the one closest to the remaining untreated, etc.
4. Select treated subjects in a random order. Set a fixed random number seed so that the matched sample is reproducible in subsequent analyses.

Results in all treated subjects being matched to a single control.

# Greedy Matching with Replacement

Matching *without* replacement means that once an untreated subject has been matched to a treated subject that untreated subject is no longer eligible for further matches to other treated subjects. As a result, each subject can be in at most one matched pair.

Now, in matching *with* replacement, we allow members of the "control" pool to be reused in the matching process.

- The process is somewhat simpler in the nearest neighbor case - just match each treated subject to the closest untreated subject.
- Because untreated subjects are recycled and thus can be included in multiple matched sets, the order in which the treated subjects are selected has no effect on the formation of matched pairs.

# Matching 1:k rather than 1:1

Here, we simply try to obtain the k best matching untreated subjects for each treated subject.

- In greedy matching, it is certainly possible that the quality of matches will drop considerably with extra matches, especially near the edges of the distribution of the propensity score.
- 1:k matching is occasionally done with replacement, but of course we still want k unique matched untreated subjects for each treated subject.

# Caliper Matching

Match subjects only if they fall within a pre-specified maximum distance (the caliper distance.)

- When using caliper matching, we usually match subjects on the logit of the propensity score using a caliper width as a proportion of the standard deviation of the logit of the propensity score.
- Caliper matching can be combined with other distance metrics (where, for example, a few specific covariates are targeted for more precise matching.)
- Matching with a caliper can be accomplished with or without replacement, and in 1:1 or 1:k settings.

# Optimal Matching

The main distinction that matters is between optimal matching approaches and nearest-neighbor (greedy) matching approaches.

- Optimal matching forms matched pairs so as to minimize the average within-pair difference in propensity scores.
- Optimal matching is rarely the first way I run an analysis (it's a bit slow, especially with large matching problems) but this problem is disappearing as smarter people and more effective computers emerge.

# Double Robust Approaches

Nothing is stopping us from using regression adjustment along with matching. It's not unusual to consider the incorporation of the linear propensity score, or an important set of prognostic covariates in a setting where we are analyzing propensity-matched subjects.

# Peter Austin's (2014) Comparison

## A comparison of 12 algorithms for matching on the propensity score

**Peter C. Austin**[a,b,c*†]

**Propensity-score matching is increasingly being used to reduce the confounding that can occur in observational studies examining the effects of treatments or interventions on outcomes. We used Monte Carlo simulations to examine the following algorithms for forming matched pairs of treated and untreated subjects: optimal matching, greedy nearest neighbor matching without replacement, and greedy nearest neighbor matching without replacement within specified caliper widths. For each of the latter two algorithms, we examined four different sub-algorithms defined by the order in which treated subjects were selected for matching to an untreated subject: lowest to highest propensity score, highest to lowest propensity score, best match first, and random order. We also examined matching with replacement. We found that (i) nearest neighbor matching induced the same balance in baseline covariates as did optimal matching; (ii) when at least some of the covariates were continuous, caliper matching tended to induce balance on baseline covariates that was at least as good as the other algorithms; (iii) caliper matching tended to result in estimates of treatment effect with less bias compared with optimal and nearest neighbor matching; (iv) optimal and nearest neighbor matching resulted in estimates of treatment effect with negligibly less variability than did caliper matching; (v) caliper matching had amongst the best performance when assessed using mean squared error; (vi) the order in which treated subjects were selected for matching had at most a modest effect on estimation; and (vii) matching with replacement did not have superior performance compared with caliper matching without replacement.** © 2013 The Authors. ***Statistics in Medicine*** **published by John Wiley & Sons, Ltd.**

**Keywords:** propensity score; matching; computer algorithms; optimal matching; Monte Carlo simulations; propensity-score matching

- You'll find this article on our Sources page.

# Austin's conclusions re: 12 Algorithms

- Larger numbers of matched pairs (from complete optimal or complete greedy matches) yields more precise estimates than smaller numbers of matched pairs (say, when a caliper is used and only some treated subjects are matched.)
- Caliper matching often yields better "balance" and less biased estimates as compared to other algorithms.
- So we have a bias - variance tradeoff in our estimation strategies, but in terms of MSE, caliper matching usually performs pretty well.
- In terms of ordering of treated subjects for greedy matching or caliper matching, random selection is competitive with other options.
- Optimal matching is pretty comparable to nearest neighbor matching with random selection order, and in fact, it's not clearly any better than that approach.

# Section 2

## Feinstein's Model for Research Architecture (expanded by Neal Dawson)

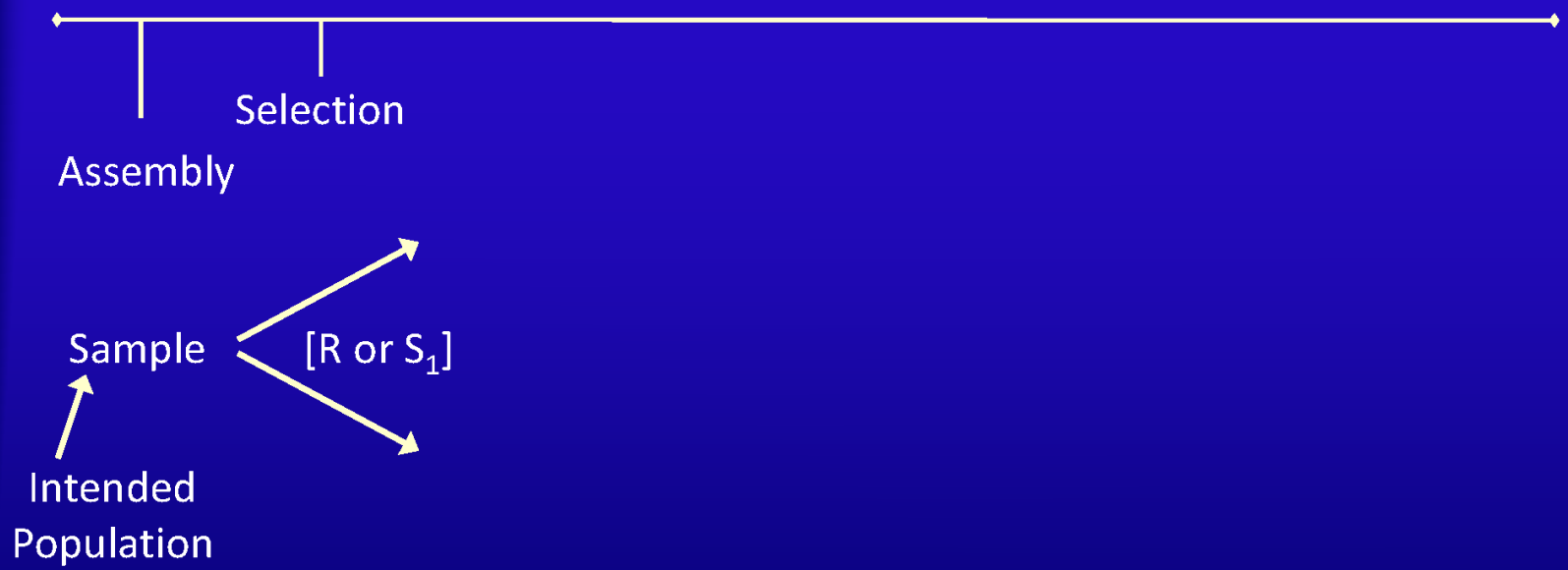
# 7 Key Aspects of Research Architecture*:
## Making Fair Judgments about Causation


- Possibility of distorted assembly – sample doesn't reflect the population to which the results will be generalized.

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

3

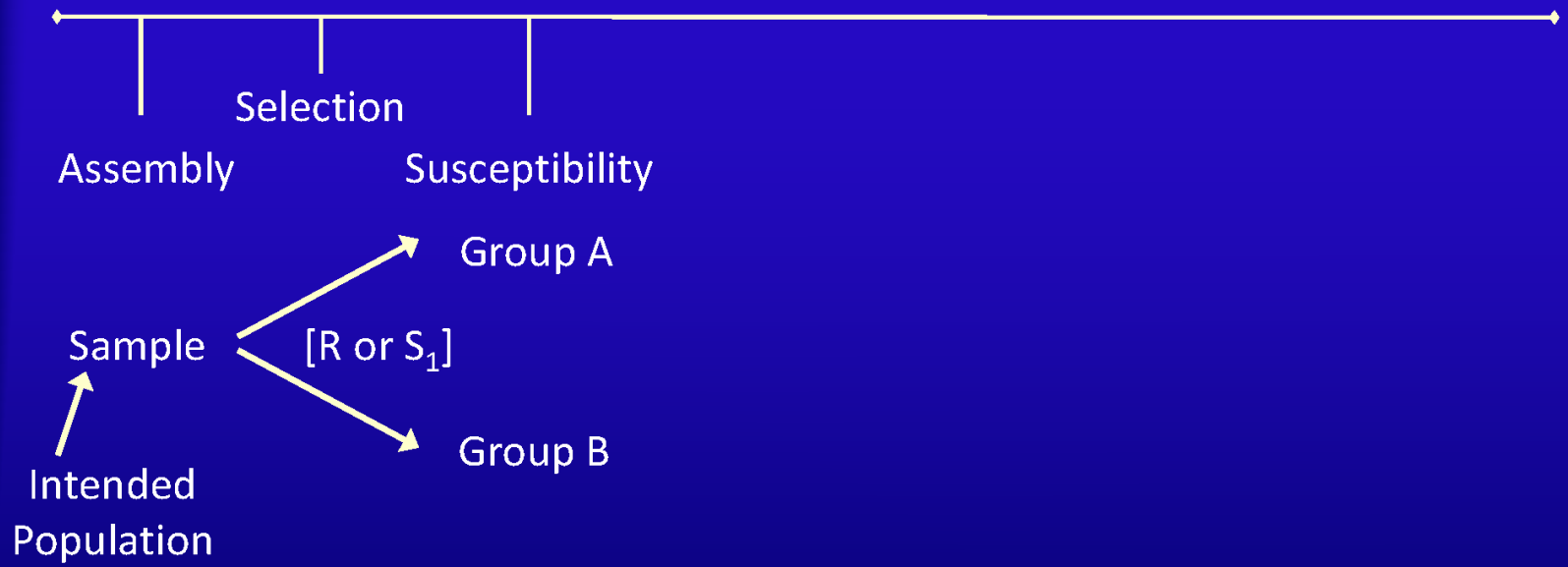
# 7 Key Aspects of Research Architecture*:
## Making Fair Judgments about Causation




- Selection Bias – who receives the exposure? Basis: (possibly unmeasured) covariates linked to outcomes? Why randomize?

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

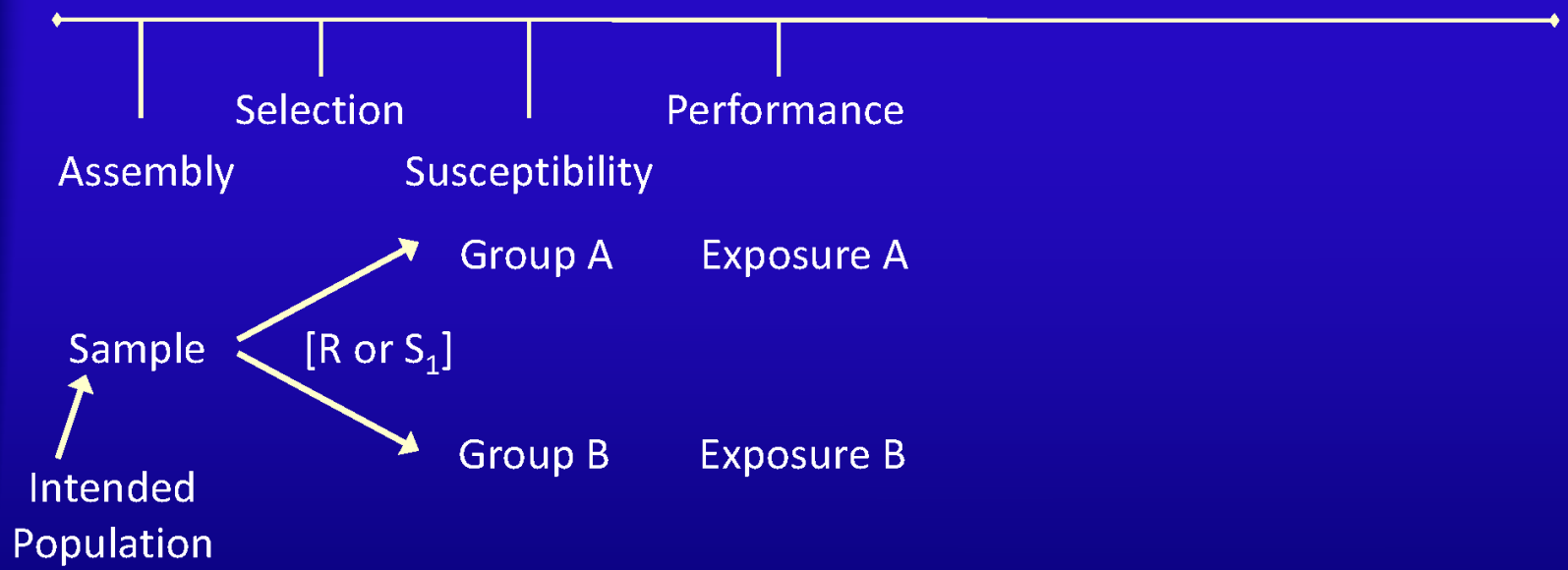
# 7 Key Aspects of Research Architecture*: Making Fair Judgments about Causation




- Are there importantly different expectations at baseline, for the eventual outcomes? Susceptibility reflects covariate differences.

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

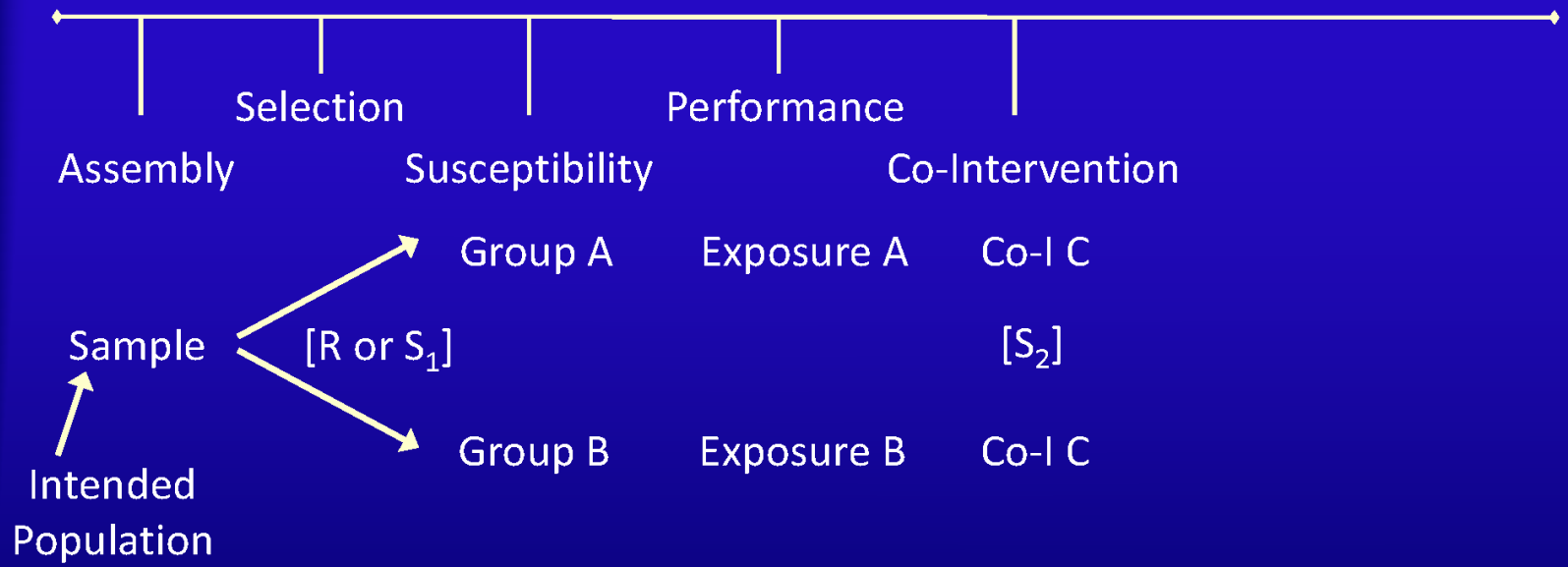
# 7 Key Aspects of Research Architecture*:
## Making Fair Judgments about Causation




- Are exposures applied with the same proficiency? How "well" do pts receive the exposures (dosage schedules, compliance)?

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

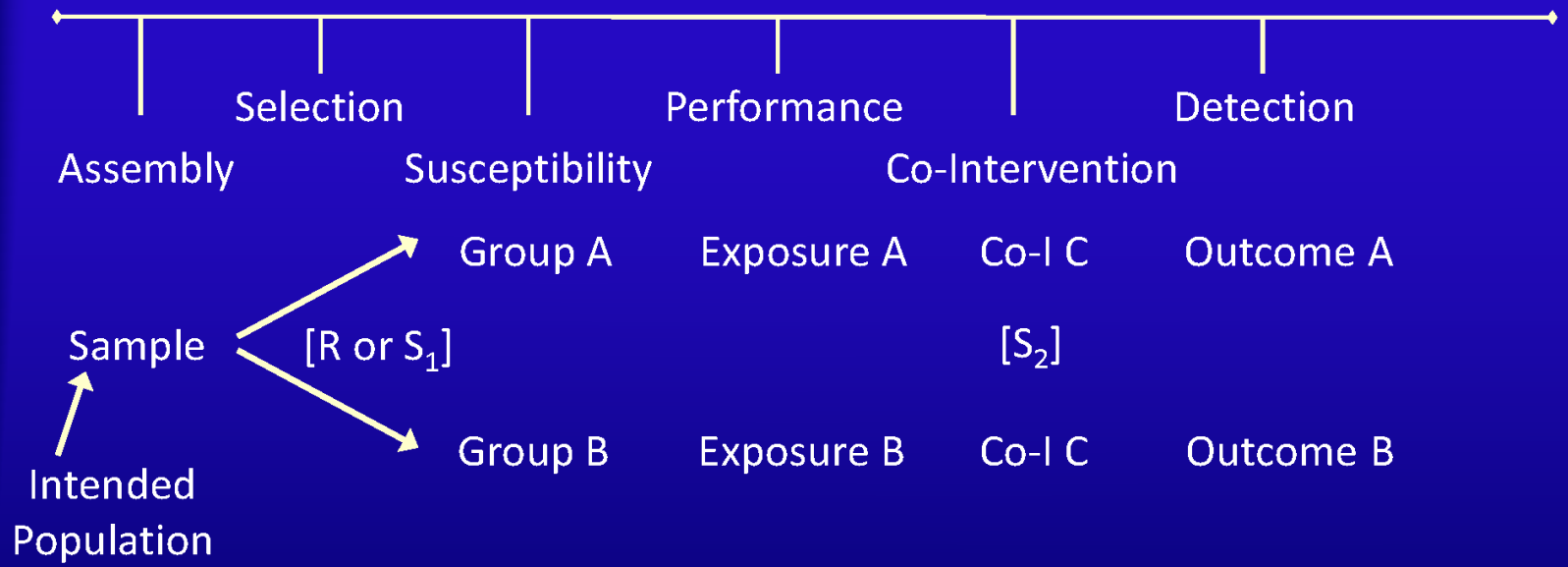
# 7 Key Aspects of Research Architecture*: Making Fair Judgments about Causation




- Additional selection opportunities – co-interventions (beyond exposure of interest) may influence likelihood of outcomes.

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

# 7 Key Aspects of Research Architecture*:
## Making Fair Judgments about Causation

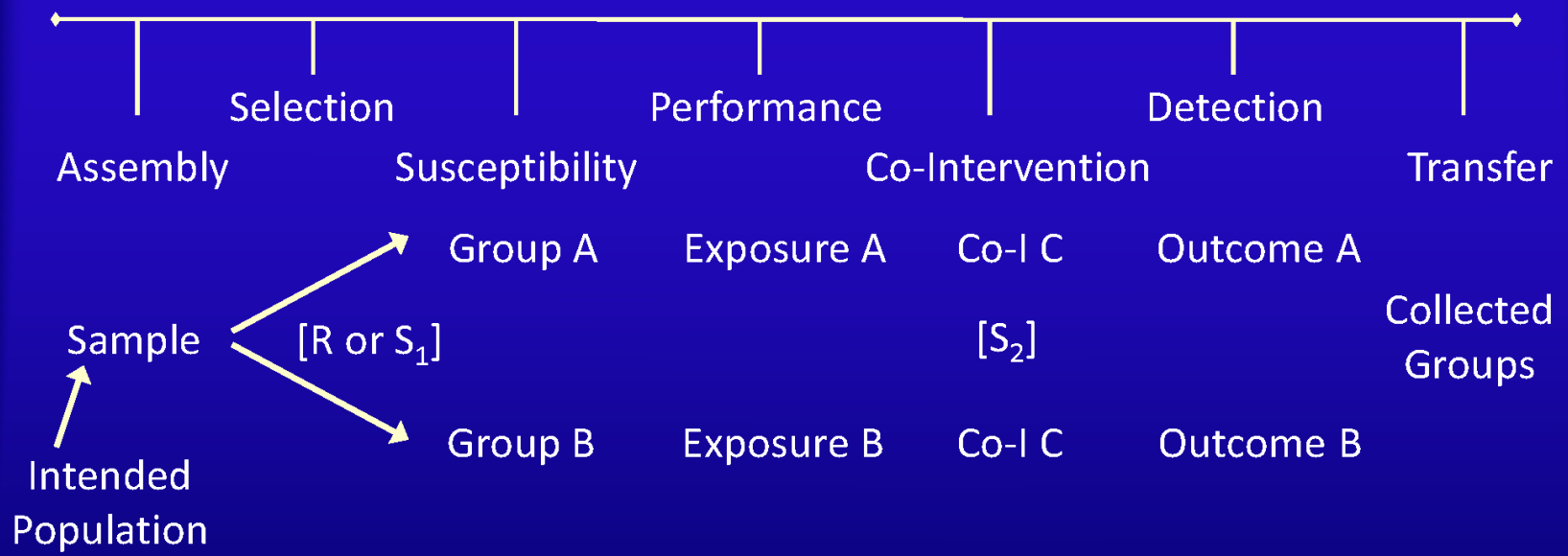



- Is process for determining outcomes applied unequally? Differences in surveillance, diagnostic testing, or interpretation?

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

8

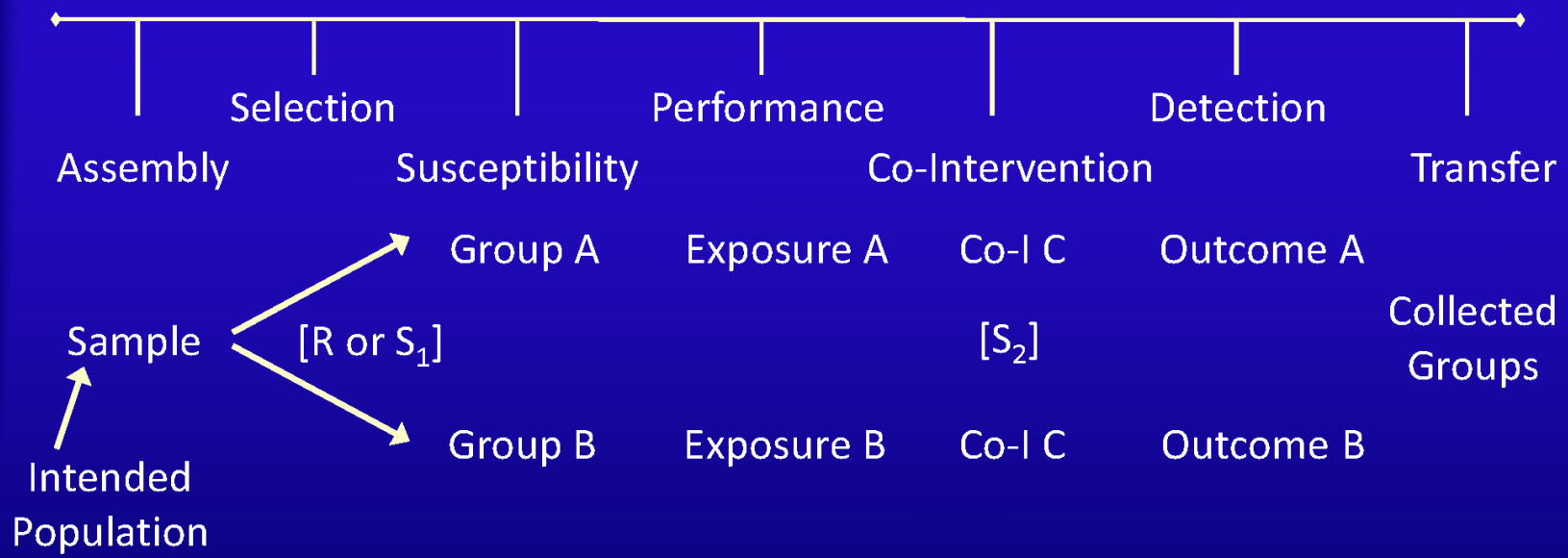
# 7 Key Aspects of Research Architecture*:
## Making Fair Judgments about Causation




- Comparison of members of original cohorts of A and B – dropouts, in-study exclusions, crossovers, dealing with missing data…

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

# 7 Key Aspects of Research Architecture*:
## Making Fair Judgments about Causation




- Goal: Comparability of groups who did and did not receive the exposure (except for the actual receipt of the exposure)

*Adapted by Neal Dawson from Alvan Feinstein's intellectual model (5 key aspects)

# Section 3

## Designing Observational Studies (Rubin 2001)

# On Designing Observational Studies

- Exert as much experimental control as possible
- Carefully consider the selection process
- Actively collect data to reveal potential biases

  *"Care in design and implementation will be rewarded with useful and clear study conclusions... Elaborate analytical methods will not salvage poor design or implementation of a study." – NAS report (quoted in Rosenbaum p. 368)*

But **HOW**?

# Rubin 2001

On Designing an Observational Study with the Propensity Score

## Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation

DONALD B. RUBIN

# Rubin 2001 Abstract

**Abstract.** Propensity score methodology can be used to help design observational studies in a way analogous to the way randomized experiments are designed: without seeing any answers involving outcome variables. The typical models used to analyze observational data (e.g., least squares regressions, difference of difference methods) involve outcomes, and so cannot be used for design in this sense. Because the propensity score is a function only of covariates, not outcomes, repeated analyses attempting to balance covariate distributions across treatment groups do not bias estimates of the treatment effect on outcome variables. This theme will the primary focus of this article: how to use the techniques of matching, subclassification and/or weighting to help design observational studies. The article also proposes a new diagnostic table to aid in this endeavor, which is especially useful when there are many covariates under consideration. The conclusion of the initial design phase may be that the treatment and control groups are too far apart to produce reliable effect estimates without heroic modeling assumptions. In such cases, it may be wisest to abandon the intended observational study, and search for a more acceptable data set where such heroic modeling assumptions are not necessary. The ideas and techniques will be illustrated using the initial design of an observational study for use in the tobacco litigation based on the NMES data set.

**Keywords:** balance, matching, subclassification

# Designing an Observational Study without access to the outcome data

- Propensity score methods can be used to help design the OS without seeing any outcomes.
  - Propensity score is a function only of covariates, not of outcomes.

## The key insight from Rubin (2001)

Repeated analyses attempting to balance covariate distributions across treatment groups **do not bias** estimates of the treatment's effect on outcome variables.

# Designing Observational Studies

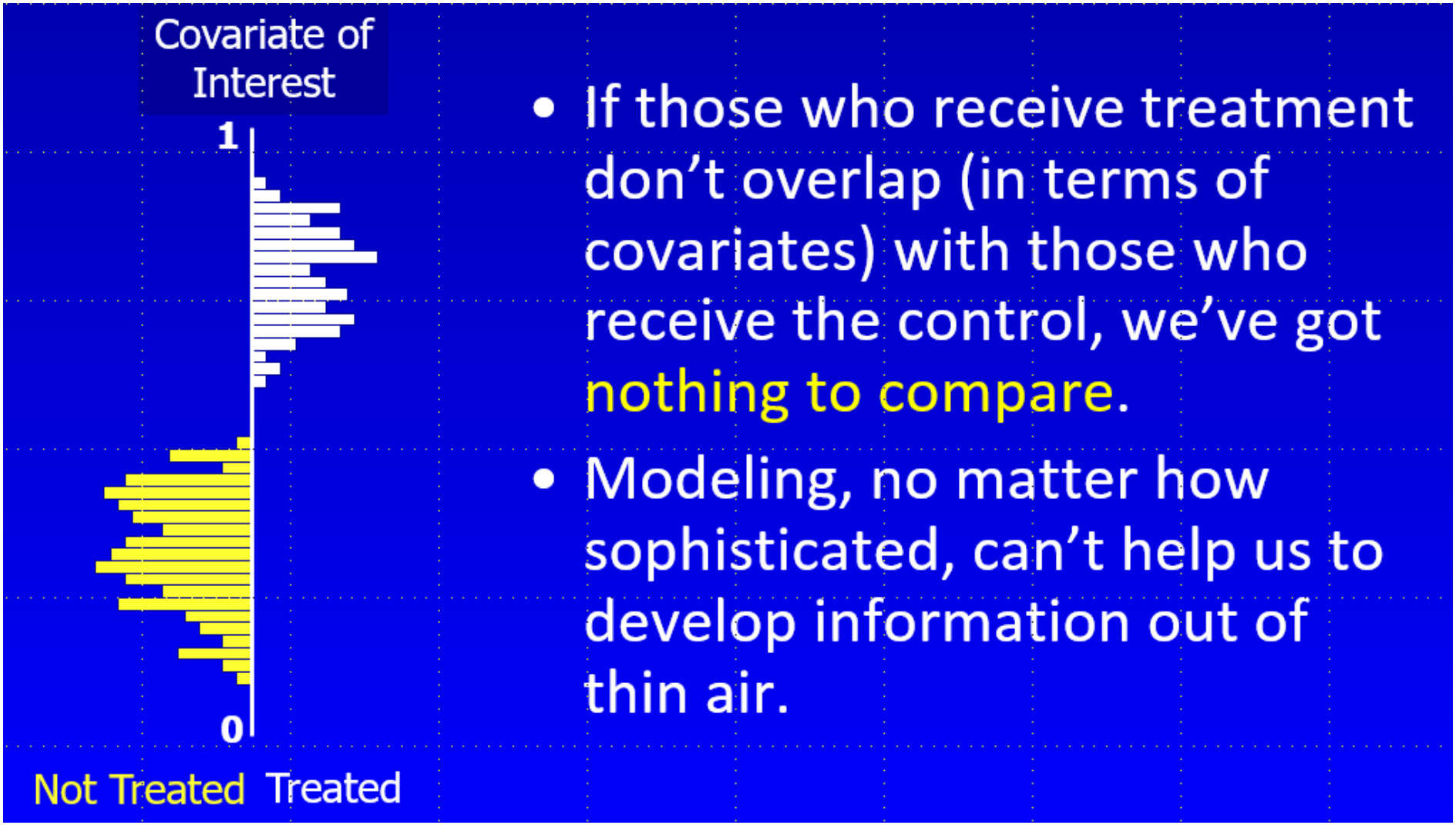
- The Importance of Covariate (PS) Overlap
- How To Check for Overlap Effectively
- Designing Like You're Doing an Experiment
- Using Matching, Subclassification and Weighting
- Propensity Scores are "Fair Game" - No Outcomes!

In order to extract information on treatment effect from an observational study, we need to be able to compare "identical" people who receive different treatments.

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

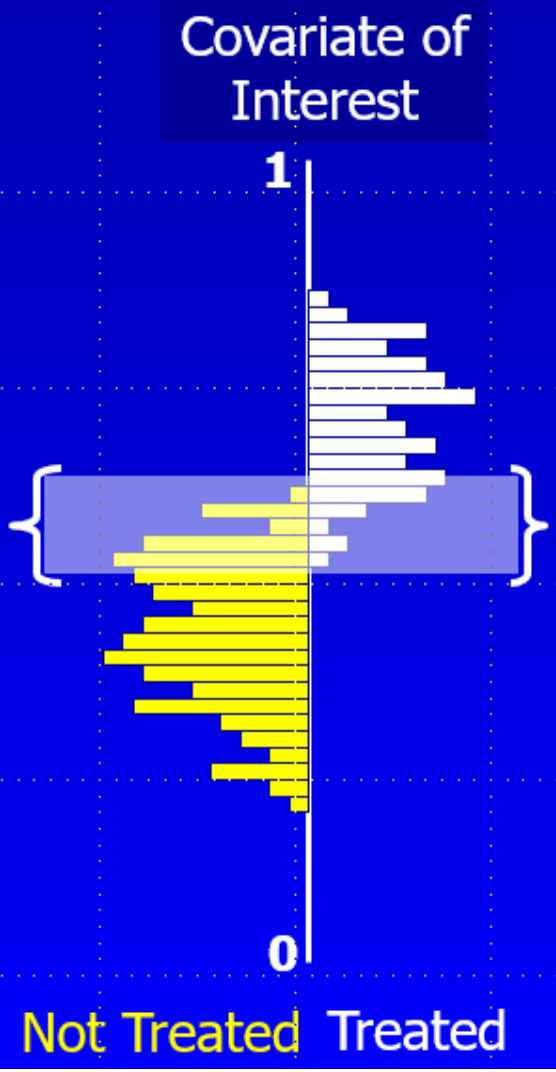
**Issue 1**: Overlap

# How much overlap in the covariates do we want?




- 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 the exposure groups overlap, but minimally?




- No help.
- The information available to infer treatment effect will reside almost entirely in the few patients who overlap.
- Need to think hard about whether useful inferences will be possible.

# Initial Phases of Ideal Study Design

Specify population, exposures/treatments, outcomes and covariates.

- Collect treatment and covariate information, and model treatment assignment with the propensity score.
- Use propensity scores (through matching, stratification, reweighting) to reduce bias.
- Check for covariate balance across the treatment groups and iterate through process.
  - If the treatment and control groups have the same distribution of propensity scores, then they have the same distribution of all observed covariates, just as in a randomized experiment.
  - Of course, propensity scores are only guaranteed to balance the observed covariates, while randomized experiments can stochastically balance unobserved, as well.

# Rich and Poor Covariate Sets

- With a rich set of covariates, adjustments for hidden covariates may be less critical.
- With less rich covariate sets, we may need to do more, say, try to find an instrument.

As Rubin mentions in the Abstract, our conclusion after the initial design stage may be that the treatment and control groups are too far apart to produce reliable effect estimates without heroic modeling assumptions.

# Techniques for Initial Observational Study Design using Propensity Scores

- Matching
- Subclassification / Stratification
- Weighting

**Goal**: Assemble groups of treated and control units such that within each group the distribution of covariates is balanced.

Allows us to attribute outcome differences to the effect of treatment vs. control.

# Why Work this Hard in the Initial Design Stage?

- Options narrow as an investigation proceeds.
- No harm, no foul.
  - Since no outcome data are available to the PS, nothing based on the PS here biases estimation of treatment effects.
- Balancing covariates / PS makes subsequent model-based adjustments more reliable.

Key point is that model adjustments can be extremely unreliable when the treatment groups are far apart on covariates. So we need to avoid that.

"Balancing" helps in terms of assessing covariance, relative risk, subsequent adjustments, etc.

# Propensity Score **Matching** in the Design of an Observational Study

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

# Propensity Score **Subclassification** in the Design of an Observational Study

- Rank all subjects by their propensity score and then create subclasses by imposing boundaries.
- Subclasses therefore have treated and control units with similar values of propensity score.
- Often use 5 subclasses of equal size should remove 90% or more of the bias due to the observed covariates in the propensity score.

# Propensity Score **Weighting** in the Design of an Observational Study

Estimate propensity scores for each subject, so that PS = prob(treatment received | covariates)

Rubin describes the ATE approach to weighting...

- Weights for treated subjects: $\frac{1}{PS}$.
- Weights for control subjects: $\frac{1}{1-PS}$

# When Can We Move On?

Three conditions which must all apply for regression adjustment to be trustworthy:

1. Difference in the means of linear propensity score [logit(PS)] in the two groups being compared must be small.
2. Ratio of variances of linear propensity scores in the two groups must be close to 1.
3. Ratio of variances of the "residuals" of the covariates after PS adjustment close to 1.

These are what I have referred to as "Rubin's Rules"...

# Three Rules (page 174, Rubin 2001)

In particular, there are three basic distributional conditions that in general practice must simultaneously obtain for regression adjustment (whether by ordinary linear regression, linear logistic regression, or linear-log regression) to be trustworthy. If any of these conditions is not satisfied, the differences between the distributions of covariates in the two groups must be regarded as substantial, and regression adjustment will be unreliable and cannot be trusted. These conditions are:

1. The difference in the means of the propensity scores in the two groups being compared must be small (e.g., the means must be less than half a standard deviation apart), unless the situation is benign in the sense that: (a) the distributions of the covariates in both groups are nearly symmetric, (b) the distributions of the covariates in both groups have nearly the same variances, and (c) the sample sizes are approximately the same.
2. The ratio of the variances of the propensity score in the two groups must be close to one (e.g., 1/2 or 2 are far too extreme).
3. The ratio of the variances of the residuals of the covariates after adjusting for the propensity score must be close to one (e.g., 1/2 or 2 are far too extreme); "residuals" precisely defined shortly.

# Assessing Balance on the *Linear* rather than *Raw* Propensity Score

- *logit*(PS) is more relevant for assessing whether linear modeling adjustments work.
- *logit*(PS) tend to have more benign (variances closer, greater symmetry) distributions.
- *logit*(PS) are more closely related to benchmarks in the literature on adjustments for covariates based on linearity assumptions.

# Putting Rubin's Rule 1 into operation

1. Difference in the means of the propensity scores in the two groups being compared.
    - Estimate propensity scores for all subjects.
    - Take logit(PS) for each subject (normalize).
    - Find SD = standard deviation of logit(PS) across all subjects (treated and control).
    - Mean logit(PS) for treated group should be within 0.5 SD of control group's mean logit(PS).
    - Often we calculate a standardized difference here.

```
rubin1.unadj <- with(toy, abs(100*(mean(linps[treated==1]) -
                                   mean(linps[treated==0]))
                                     sd(linps)))
```

# Putting Rubin's Rule 2 into operation

2. Variance ratio of propensity scores in the two groups being compared should be close to 1.
    - Estimate propensity scores for all subjects.
    - Take logit(PS) for each subject (normalize).
    - Find variance of logit(PS) across treated subjects, and divide it by the variance of logit(PS) across control subjects.
    - Variance ratio should be close to 1. Ratios of 0.5 and 2.0 are far too extreme: we often try for (4/5, 5/4).

```
rubin2.unadj <-with(toy, var(linps[treated==1]) /
                         var(linps[treated==0]))
```

# Putting Rubin's Rule 3 into operation

3. Variance ratio of "residuals" close to 1.
   - Estimate propensity scores for all subjects.
   - For each covariate, regress the original value of the covariate for each subject on logit(PS) and take the residual of this regression.
   - For each covariate, divide variance of the residuals within treatment group by variance of the residuals within control group.
   - For each covariate, this variance ratio should also be close to 1 (2 or 0.5 are, again, far too extreme).

# rubin3 function built for the toy example

```
## General function rubin3 to help calculate Rubin's Rule 3
rubin3 <- function(data, covlist, linps) {
  covlist2 <- as.matrix(covlist)
  res <- NA
  for(i in 1:ncol(covlist2)) {
    cov <- as.numeric(covlist2[,i])
    num <- var(resid(lm(cov ~ data$linps))[data$treated==1])
    den <- var(resid(lm(cov ~ data$linps))[data$treated==0])
    res[i] <- round(num/den, 3)
  }
  names(res) <- names(covlist)
  print(res)
}
```

# National Medical Examination Survey

- Large nationally representative data base of nearly 30,000 adults, calendar year 1987
- Modern related efforts are folded into NHANES

**Goal**: objective observational study on the causal effects of smoking and the effect of the tobacco companies' alleged misconduct

# NMES Covariates for Smoking Study

- Age, Sex, Race, Marital Status, Education, etc.
- Detailed smoking information
  - Classification of subjects as never smokers, former smokers and current smokers
  - Further classifications possible by length and density of smoking behaviors
  - Also can look at years since quitting for former smokers

# NMES Objects of Inference

- Smoking Attributable Fractions
- Conduct Attributable Fractions
- Relative Expenditure Risks

All based on comparisons of specific health-related expenditures (or disease rates)

Comparisons of smokers with "never smokers" who have same covariate values, as a function of dosage and covariates

# Rubin's Main Example

**Design Goal**: Create samples of smokers and never smokers in NMES with the same multivariate distribution of covariates.

- Males and Females treated separately.
- We'll focus first on Male "Current Smokers" vs. Male "Never Smokers"
  - 3510 Male "Current Smokers" in the pool
  - 4297 Male "Never Smokers"" as controls
- Fit propensity for "current smoker" to these people, via logistic regression with sampling weights

Separate models were built for "former vs. never (Males)" and the two analogous comparisons of Females.

# Propensity Model: 146 Covariates

| Variables Used in Propensity Model | Description |
|---|---|
| Seatbelt | 5 levels of reported seat belt use |
| Arthritis | Whether reported suffering from arthritis |
| Census Division | 9 census regions |
| Champ Insurance | Whether have military insurance |
| Diabetes | Doctor ever told having diabetes |
| Down time | 6 levels of reported emotional down time |
| Dump time | 6 levels of reported in the dumps time |
| Employment | Indicating employment status each quarter |
| English | English is a primary language |
| Retirement | Indicator for retirement status |
| Number of Friends | 7 levels measuring the number of friends |
| Membership in Clubs | 6 levels measuring memberships in clubs |
| Education | Completed years of education |
| HMO coverage | Indicating HMO coverage each quarter |

# Propensity Model: 146 Covariates

| | |
|---|---|
| High blood pressure | Doctor ever told having high blood pressure |
| Industry Code | 14 Industry codes |
| Age | Age of the respondent |
| Labor Union | Indicator for a member of labor union |
| Log Height | Natural Logarithm of height |
| Log Weight | Natural Logarithm of weight |
| Marital Status | Marital status in each quarter |
| Medicaid | On medicaid (each quarter) |
| Medicare | On medicare (each quarter) |
| Occupation | Occupation code (13 levels) |
| Public Assistance | Other public assistance program (each quarter) |
| Friends over | Frequency of having friends over (7 levels) |
| Physical Activity | Indicator variable for physically active |
| Population density | 3 levels |
| Poverty Status | 6 levels |
| Pregnant 1987 | Pregnancy status in 1987 (women) |
| Private Insurance | Other private insurance (each quarter) |
| Race | 4 levels |

# Propensity Model: 146 Covariates

| | |
|---|---|
| Race | 4 levels |
| Rated Health | 5-point self rating of health status |
| Home ownership | Indicator for owning home |
| Rheumatism | Indicator for suffering from rheumatism |
| Share Life | Indicator variable for having somebody to share their life |
| Region | 4 levels of region of the country |
| MSA | 4 levels indicating types of metropolitan statistical area |
| Risk | General risk taking attitude (5 levels) |
| Uninsured | Indicator for lack insurance (each quarter) |
| Veteran | Indicator for veteran status |
| Incapler | Survey weight in NMES database |
| Agesq | Age*Age |
| Educat.sq | Education*Education |
| Age_wt | Age*Logwt |
| Age_educt | Age*Education |
| Age_ht | Age*Loght |
| Educat_wt | Education*Logwt |

# Propensity Model: 146 Covariates

| Variables Used in Propensity Model | Description |
|---|---|
| Educat_ht | Education*Loght |
| Loght_logwt | Loght*Logwt |
| Loghtsq | Loght*Loght |
| Logwtsq | Logwt*Logwt |

# Assessing Overlap Step 1: Looking for Mean Bias

Bias B = standardized difference in the means of logit(propensity scores) between current smokers and never smokers for males

- We want the bias in the propensity score to be small, no greater than 0.50 in absolute value.
- Here, mean propensity score Bias B = 1.09 (109%)
- In fact standardized difference $> 0.5$ (50%) for many of the individual covariates, as well.

# Assessing Overlap Step 2: Comparing Variances

Ratio R = ratio of the variances of logit(propensity scores) between current smokers and never smokers for males.

- We want the variances to be homogeneous, so the ratio should be close to 1 (1/2 and 2 are far too extreme).
- Here, variance ratio for logit(PS) is R = 1.00
- Could look at ratio of individual covariate variances, also. (In fact, `MatchBalance` does this.)

# Assessing Overlap Step 3: Comparing Residuals

Regress each covariate on logit(PS) and look at ratio of variances of residuals for current smokers to variance of residuals for never smokers within the male population.

- Here, we get a separate result for each of the 146 covariates. We want results near 1.00
  - 57% of the covariates had their residual ratio between 4/5 and 5/4
  - 5% of covariates had their residual ratio below 1/2, or above 2

# Excerpt from Rubin's Table 2 (page 179)

*Table 2.* Estimated propensity scores on the logit scale for "smokers" versus never smokers in full NMES

| | | | Percent of covariates with specified variance ratio orthogonal to the propensity score | | | | |
|---|---|---|---|---|---|---|---|
| Treated Group | $B$ | $R$ | $\leq 1/2$ | $>1/2$ and $\leq 4/5$ | $>4/5$ and $\leq 5/4$ | $>5/4$ and $\leq 2$ | $>2$ |
| Male Current $N=3{,}510$ | 1.09 | 1.00 | 3 | 9 | 57 | 26 | 5 |

$B$ = Bias, $R$ = Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.

**Interpretation**

> *"... [A]ny linear (or part linear) regression model cannot be said to adjust reliably for these covariates, even if they were perfectly normally distributed. ... B [is] greater than 1/2, and many of the value of R for the residuals of the covariates are outside the range (4/5, 5/4)."*

# Similar Results for the Other Study Comparisons

*Table 2.* Estimated propensity scores on the logit scale for "smokers" versus never smokers in full NMES

| | | | Percent of covariates with specified variance ratio orthogonal to the propensity score | | | | |
|---|---|---|---|---|---|---|---|
| Treated Group | $B$ | $R$ | $\leq 1/2$ | $>1/2$ and $\leq 4/5$ | $>4/5$ and $\leq 5/4$ | $>5/4$ and $\leq 2$ | $>2$ |
| Male Current $N=3{,}510$ | 1.09 | 1.00 | 3 | 9 | 57 | 26 | 5 |
| Male Former $N=3{,}384$ | 1.06 | 0.82 | 2 | 15 | 61 | 15 | 7 |
| Female Current $N=3{,}434$ | 1.03 | 0.85 | 1 | 15 | 59 | 23 | 2 |
| Female Former $N=2{,}657$ | 0.65 | 1.02 | 5 | 7 | 85 | 7 | 5 |

$B=$ Bias, $R=$ Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.

All four comparisons indicate the need for propensity score adjustments.

# Mahalanobis Matching within PS Calipers

For the 3510 male current smokers, 3510 "matching" male never smokers were chosen from the pool of 4297 male never smokers.

- Method: Mahalanobis metric matching within propensity score calipers ($\pm$ 0.2 of the standard deviation of linear propensity scores)
  - Mahalanobis distance variables were: age, education, body mass index, and sampling weight.
  - Some of these are survey results, mostly (but not completely) in categories.
- In this case, there were no current smoker Males that could not be matched within the PS calipers to never smoker Males.
  - What if there had been a treated subject whose propensity score was not "matchable"?
  - What if the "donor pool" of never smokers had been empty for one of the current smokers?

# Impact of Matching on Overlap

**Male Current Smokers vs. Male Never Smokers**

| Scenario | Bias, B | Variance Ratio, R |
|---|---|---|
| Before Matching | 1.09 | 1.00 |
| After Matching | 0.08 | 1.16 |

Residual Variance Ratios (% in range)

| Range | Before Match | After Match |
|---|---|---|
| $\leq 0.5$ | 3 | 1 |
| $(\frac{1}{2}, \frac{4}{5}]$ | 9 | 3 |
| $(\frac{4}{5}, \frac{5}{4}]$ | **57** | **90** |
| $(\frac{5}{4}, 2]$ | 26 | 6 |
| $> 2$ | 5 | 0 |

# Matching's Impact on Overlap

**Male Former Smokers vs. Male Never Smokers**

| Scenario | B | R | Res. VR in $(\frac{4}{5}, \frac{5}{4}]$ |
|---|---|---|---|
| Before Match | 1.06 | 0.82 | 61% of covariates |
| After Match | 0.04 | 0.99 | 94% |

# Female Comparisons re: Matching

Female **Current** Smokers vs. Female Never Smokers

| Scenario | B | R | Res. VR in $(\frac{4}{5}, \frac{5}{4}]$ |
|---|---|---|---|
| Before Match | 1.03 | 0.85 | 59% of covariates |
| After Match | 0.04 | 0.94 | 93% |

Female **Former** Smokers vs. Female Never Smokers

| Scenario | B | R | Res. VR in $(\frac{4}{5}, \frac{5}{4}]$ |
|---|---|---|---|
| Before Match | 0.65 | 1.02 | 85% |
| After Match | 0.06 | 1.02 | 91% |

# Re-estimating PS using Matched Subjects Only

Original propensity score estimate used all of the subjects, including those subjects who wound up being unused controls, once we matched.

- Here, they are no longer concerned with unmatched "never smokers" so they re-estimate the propensity score using only the matched samples, then look at the remaining covariate imbalance.

| Group | B | R | Res. VR in $(\frac{4}{5}, \frac{5}{4}]$ |
|---|---|---|---|
| Male, Current | 0.39 | 1.33 | 88% |
| Male, Former | 0.32 | 1.33 | 95% |
| Female, Current | 0.35 | 1.18 | 92% |
| Female, Former | 0.31 | 1.09 | 91% |

Looks better. Suppose we are still not satisfied, though.

# Subclassification of Matched Samples

Suppose we are still not satisfied...

Create two equal-size (weighted) subclasses, low and high on the linear PS.

- Treated and Control subjects with low PS are to be compared to each other.
- Treated and Control subjects with high PS are to be compared to each other.
- Weighted average of two comparisons yields the result.

# Subclassification as Re-Weighting

- For the treated subjects, the new weights implied by this subclassification are the total (weighted) number of treated and controls in that subclass, divided by the total (weighted) number of treated subjects.
- For the control subjects, weights are the subclass total of treated & controls divided by subclass controls.

Leads to a weighted PS analysis that reflects the additional balance due to subclassification.

- The same idea for weighting works no matter how many subclasses
  - One subclass is what we've had - no subclassification adjustment, just matching.
  - We'll also look at the impact of incorporating 2, 4, 6, 8, or 10 subclasses after matching...

# Current vs. Never Smoking Males: Overlap

**Matching + Post-Matching Subclassification**

| Subclasses | B | R | Res. VR in $(\frac{4}{5}, \frac{5}{4}]$ |
|---|---|---|---|
| 1 | 0.39 | 1.33 | 88% |
| 2 | 0.18 | 1.36 | 98% |
| 4 | 0.10 | 1.25 | 99% |
| 6 | 0.09 | 1.30 | 100% |
| 8 | 0.08 | 1.16 | 100% |
| 10 | 0.07 | 1.12 | 100% |

# How Far Can We Go?

We can obtain dramatic reduction in initial bias through this sort of subclassification, and we can carefully pick out just how many subclasses will be most helpful in getting the job done.

We can even do Weighted Propensity Score Analysis (using infinitely many subclasses)

- Form ATE weights directly from the estimated propensity score without subclassification.
  - Weight for treated subject: inverse of his/her propensity score (times his/her NMES weight)
  - Weight for control subject: inverse of 1 minus his/her propensity score (times NMES weight)
  - Caveat: Can get unrealistically extreme weights when estimated PS is near zero or one.

# Current vs. Never Smoking Males: Overlap

| Analysis | B | R | Res. VR in $(\frac{4}{5}, \frac{5}{4}]$ |
|---|---|---|---|
| Full Sample | 1.09 | 1.00 | 57% |
| Match full PS | 0.08 | 1.16 | 90% |
| Match new PS | 0.39 | 1.33 | 88% |
| Match, then 2 subclasses | 0.18 | 1.36 | 98% |
| 4 subclasses | 0.10 | 1.25 | 99% |
| 6 subclasses | 0.09 | 1.30 | 100% |
| 8 subclasses | 0.08 | 1.16 | 100% |
| 10 subclasses | 0.07 | 1.12 | 100% |
| Match, then Weight | 0.03 | 1.19 | 100% |

# Why Work this Hard?

- If substantial balance in covariates is obtained in this initial design stage, the exact form of the modeling adjustment is not critical.
- Similar treated and control covariate distributions implies only limited model-based sensitivity.

Why doesn't this introduce a bias for our eventual conclusions and analytic results?

# Why can we get away with this?

- We're not affecting our conclusions in a biased way, because we don't look at outcomes here.

# Why can we get away with this?

- We're not affecting our conclusions in a biased way, because we don't look at outcomes here.
- In fact, I've yet to specify the outcomes.

# NMES Outcomes for Smoking Study

- Health-care expenditures of various types
- Occurrence of various smoking-related diseases

Remember, these outcomes are never seen during the design process.

# Section 4

## Rosenbaum and Your Projects

# Rosenbaum, Chapter 5

- Focus on a particular "break" in what we've come to expect. (Simpson's Paradox, and direct adjustment.)
- "Randomization on the basis of a covariate" will be a useful notion for us in the context of developing inferences using propensity scores.
    - Estimating causal effects will require us to describe our study as analogous to a randomized study.
- Definition of the propensity score, and the key idea of combining strata with the same probability of treatment.
- Distinction between observed and hidden selection bias.

# Rosenbaum, Chapter 6

- Natural experiments are delightful things when they appear.
- Example from the 2010 Chilean earthquake drives much of what we'll see.
- What makes an observational study "reasonably compelling"
  - This is something I'd like to talk about today in the context of your Project 1 proposals.

# Project 1 Proposals

- Population (How as the sample selected?)
- Outcome (or response)
- Treatment (or exposure)
- Covariates

# Reminder: Lab 3 due March 1 at 7 PM

1. Build unadjusted estimates of the impact of an intervention on two outcomes.
2. Fit a propensity model.
3. Evaluate Rubin's Rules 1 and 2.
4. Use direct adjustment for the propensity score to obtain a new intervention effect estimate for each outcome.
5. Do 1:1 propenisty matching without replacement (includes multiple sub-parts.)
6. Compare your estimates.

# Next Time

1. Analyses of the SUPPORT / Right Heart Catheterization Study
2. Walking through a paper: the Kuo et al 2020 study