500 Class 05

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

2023-02-16

Today's Agenda

- The `toy` and `lindner` examples
- The `dm2200` example: Executing Matching in R
  - Using `Matching` vs. `MatchIt`
- Matching with (or without) a Propensity Score
  - Matching with and without replacement
  - Greedy vs. Caliper vs. Optimal Matches
- What I Hope You've Gotten So Far from Rosenbaum 1-4
- Feinstein's Model for Research Architecture
- A Few Words on Extensions to Matching
  - Constrained and "Almost Exact" matching
  - "Fine Balance" in matching
  - Full Matching and Hansen's (2004) Example

Section 1

Our R Examples

The `toy` example

The `toy` example presents methods for doing 1:1 greedy matching without replacement using the `Match` function from the `Matching` package, and for evaluating the balance before and after matching with `cobalt` and with an alternative strategy for obtaining Love plots.

- The example uses 3 Rules I attribute to Rubin (2001) for determining when a sample comparison shows sufficient balance to allow for a reasonable regression model for the outcome.
  - **Please read** Rubin (2001) in advance of Class 6, which will mostly be about that example.
- What to do in terms of a sensitivity analysis is discussed in the final section of that example, and we'll get to that later on.

The `lindner` example

The `lindner` example also does 1:1 greedy matching on the linear propensity score, in this case however there are more treated subjects than controls, so the default greedy approach taken is to not match all treated subjects, since you run out of controls before you get to them all.

- 1:1 nearest neighbor matches are demonstrated, first without replacement (in Task 4) and then **with** replacement (in Task 6).
- The "with replacement" matching has some appealing features, not least of which being that now we can realistically create ATT estimates (since we match all treated subjects to a control, repeating some controls.)
  - The match quality (in terms of a Love plot) is much better with replacement in this case.
  - This does come with a cost, though. Most subjects are matched just a few times, but some are included more than 50 times!

The dm2200 example

This example is solely about matching, and not things like subclassification, weighting and so on. The data are simulated, again, and I'm not really focusing on the outcome models, but rather just on demonstrating how to do the matching, and how to evaluate the quality of the covariate balance.

dm2200: First Four Matches

We demonstrate the use of both the `Matching` package and the `MatchIt` package. First, using `Matching`, which is what `toy` and `lindner` use, we show:

1. 1:1 nearest neighbor matching without replacement
2. 1:2 nearest neighbor matching without replacement
3. 1:3 nearest neighbor matching with replacement
4. 1:1 nearest neighbor matching without replacement within a caliper on the (linear) propensity score

dm2200: Four Additional Matches

Using the `MatchIt` package, we demonstate:

1. 1:1 nearest neighbor matching without replacement
2. 1:1 nearest neighbor matching without replacement within a caliper on the (linear) propensity score
3. 1:1 optimal matching without replacement
4. 1:2 optimal matching without replacement

There are multiple other packages in the world for propensity matching, and `cobalt` describes and supports several of them, but these are the two I have most often used.

- `MatchIt` has some features I really like, in particular, it's easier to work with it in `cobalt`, I think, and it also has one very annoying feature, in that it's hard to get the matched data set from it.

Section 2

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 3

Rosenbaum's Observation and Experiment

Chapter 1: A Randomized Trial

- Key Example: emergency treatment of septic shock
- What makes a causal question?
  - A comparison of **potential outcomes** after alternative **treatments**
- Why is causal inference difficult?
  - Because each subject receives only one of the treatments, not both.
- What distinguishes covariates from outcomes?
  - Covariates are determined prior to treatment assignment (and thus do not change on the basis of the subject's treatment assignment.)
- Why is randomization useful?
  - Many reasons, but especially because it balances covariates, both things we observe and things we don't observe.

The Road Not Taken, by Robert Frost

The Road Not Taken

BY ROBERT FROST

Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference.

Chapter 2 Structure

- Key terms defined/discussed include: Population, Covariates, Treatment Assignments, and Effects Caused by Treatments
  - You should distinguish Potential Outcomes from Observed Responses
- Average Effects can be a little subtle, and their role in thinking about covariate balance in randomized trials is described a bit.
  - Average Causal Effects are derived from asking questions about what would have happened if everyone got treatment 1? Treatment 2? As mentioned, this is a bit of a missing data problem
- Importance of randomization
  - With randomized treatment assignment, we can estimate average causal effects in the same way that we could make a statistical inference from a random sample to its associated population.

Chapter 3 Causal Inference in Randomized Experiments

- How do we test whether no effect is plausible?
  - A uniformity trial (where everyone is treated in the same way) is a helpful way of thinking about how you'd assess this
- Randomization is really making a random selection from a group of possible treatment assignments
- The Logic and Mechanics of Hypothesis Tests of No Treatment Effect
  - If the treatment effect is zero, would we ever see data like these?
  - P values, significance levels, Rejecting and Retaining the Null
- How large is the treatment effect?
  - Requires some assumptions, even in randomized trials

Chapter 4 Irrationality and Polio

- Chapter 4 isn't something that the following chapters depend on, but it opens us up to thinking about how we present/communicate results of our studies.
- Can someone express an irrational preference?
  - How could we use an experiment to demonstrate this?
- The Salk Vaccine Randomized Trial and its Parallel Observational Study
  - Why was the randomized trial Plan B?
  - How would / are we doing this differently in response to COVID-19?
- How might self-selection play a role in our evaluation of public health actions?

Chapters 5 and 6

- Chapter 5 (Between Observational Studies and Experiments) provides important foundations for the remaining chapters.
- Chapter 6 is about natural experiments, delightful things that are hard to find in practice.

Section 4

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

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

Assembly

Sample

Intended Population

- 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)

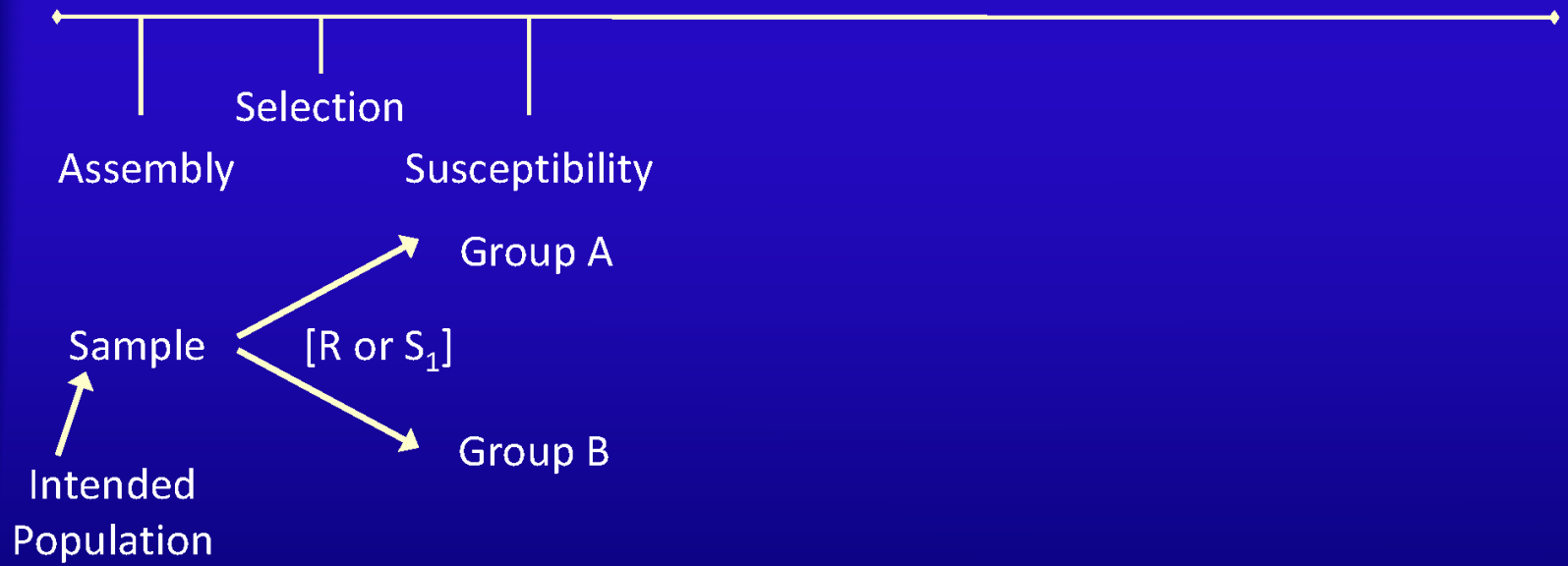
3

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)

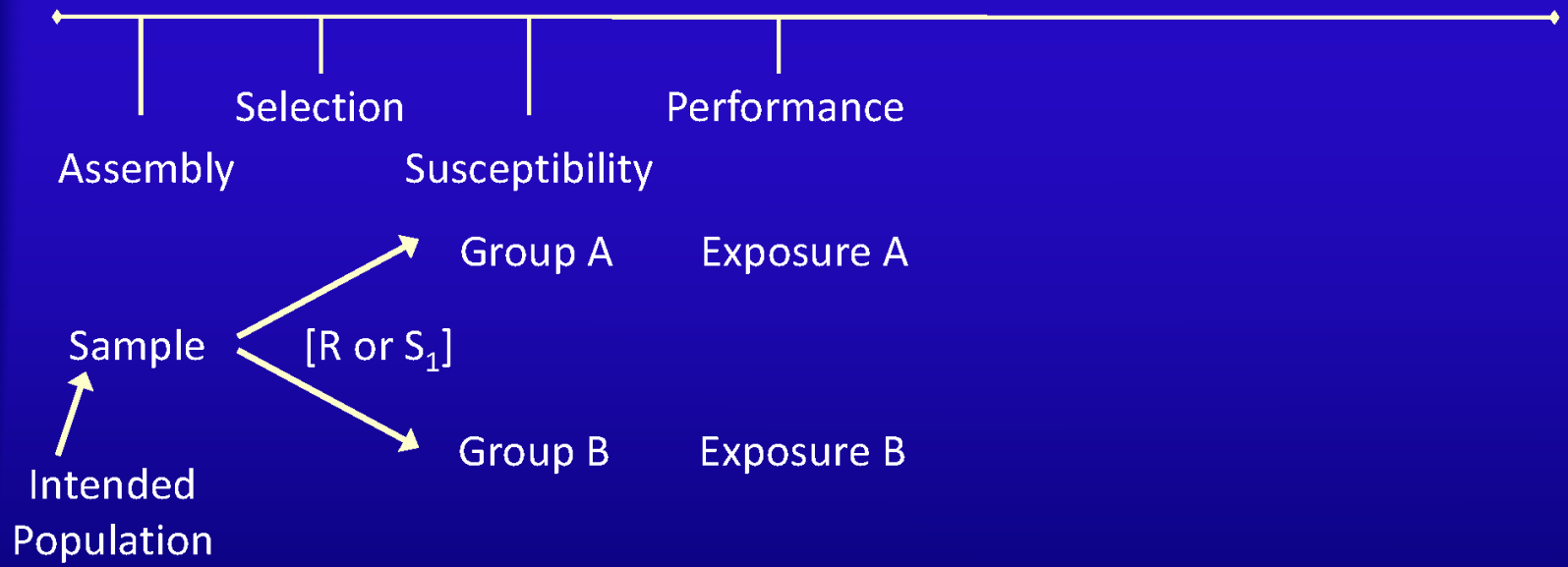
7 Key Aspects of Research Architecture*:
Making Fair Judgments about Causation



- 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)

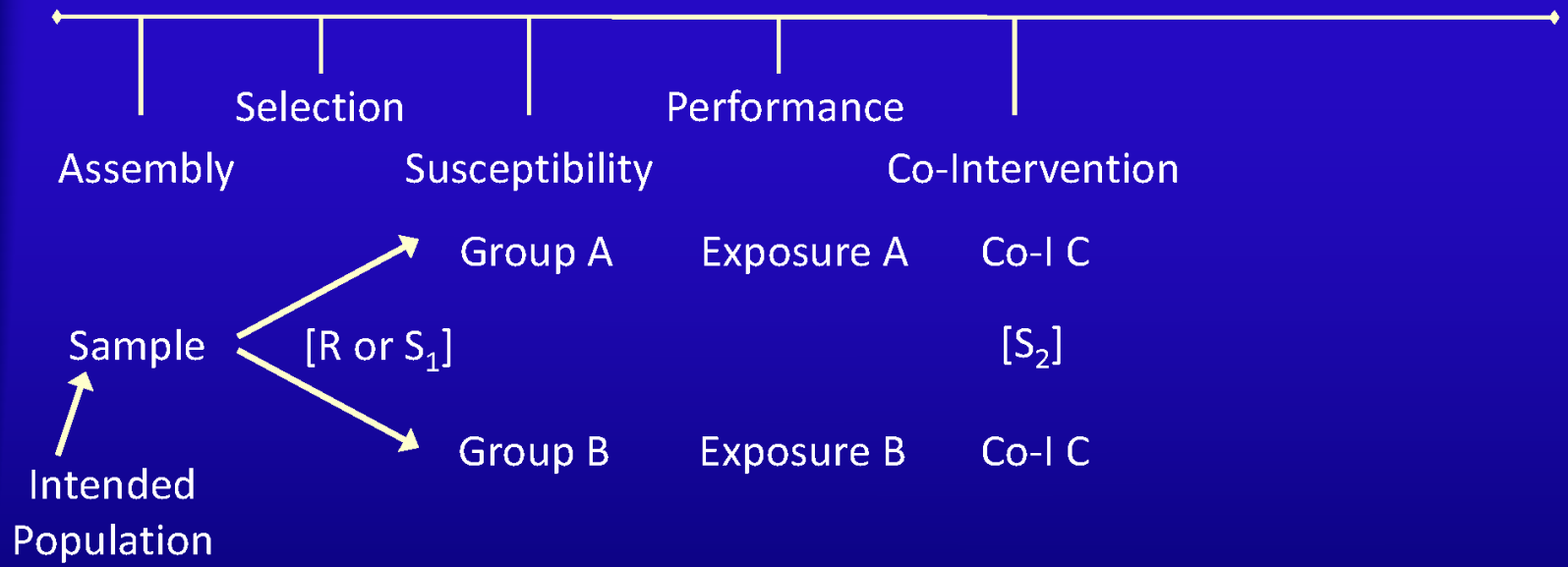
7 Key Aspects of Research Architecture*: Making Fair Judgments about Causation



- 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

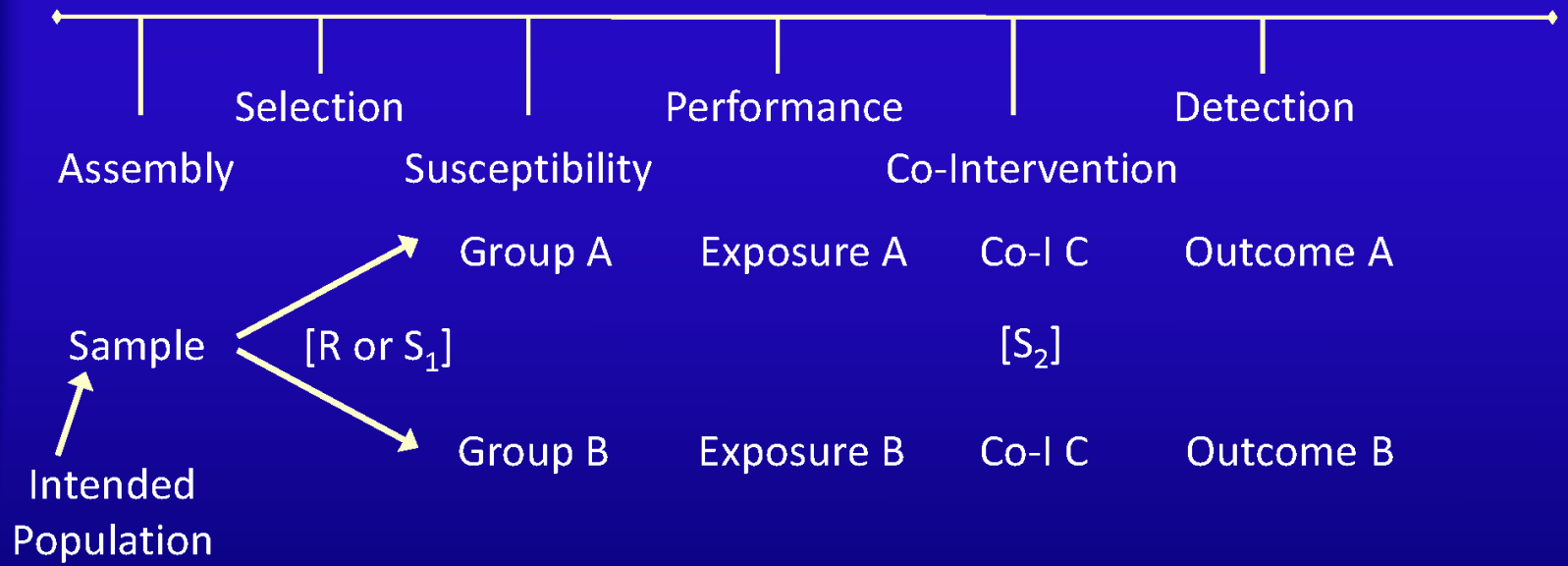


- Additional selection opportunities – co-interventions (beyond exposure of interest) may influence likelihood of outcomes.

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

7

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

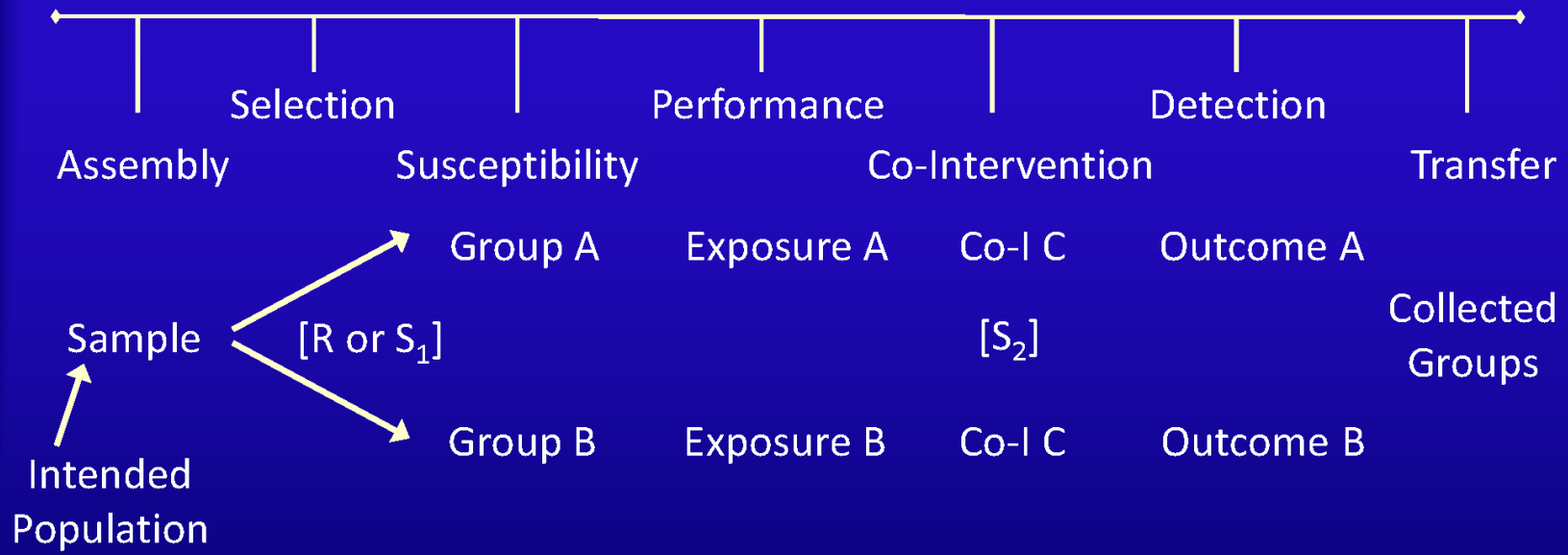


- Is process for determining outcomes applied unequally? Differences in surveillance, diagnostic testing, or interpretation?

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

8

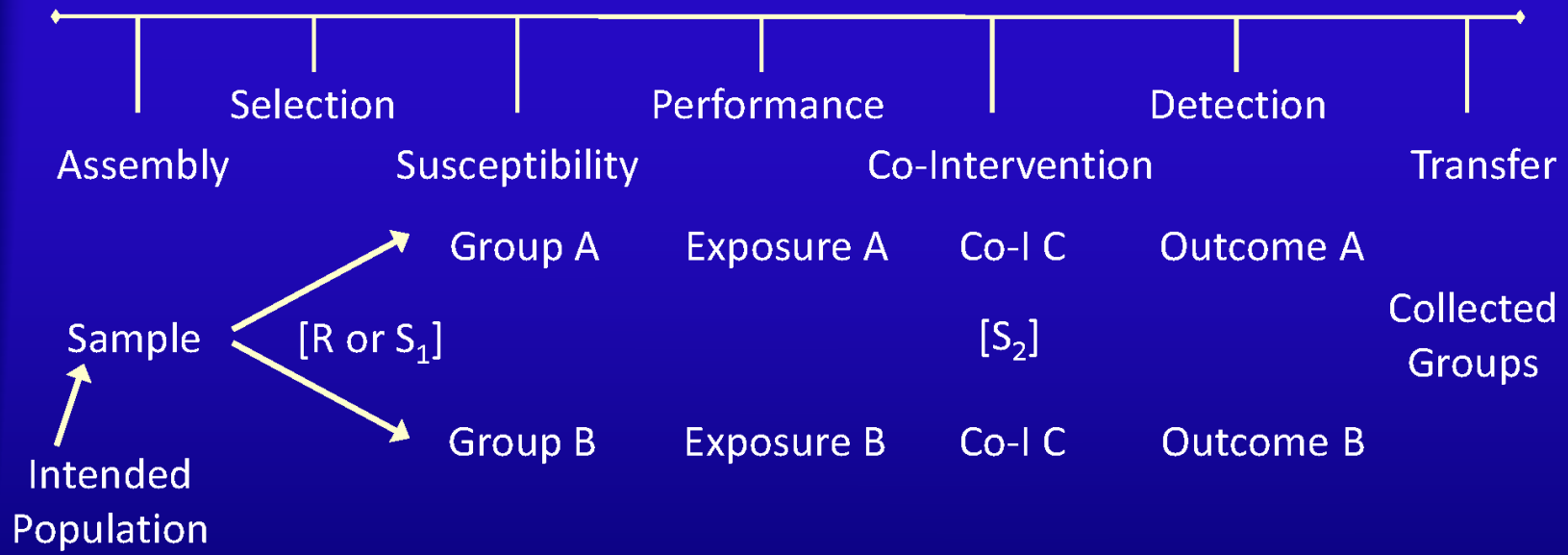
7 Key Aspects of Research Architecture*:
Making Fair Judgments about Causation



- 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 5

Some Extensions to Propensity Matching

Is Regression Adjustment Unnecessary?

- Matching and stratification are old and trusted methods of adjustment for observational studies, but the difficulty of implementing them led earlier practitioners to prefer regression.
- Modern extensions to matching methods let us perform optimal matches, full matches and optimal full matches, and to control imbalance (or at least reduce bias reduction) in ways that have become attainable only in recent years.

Good references include Rosenbaum (2010) and Hansen (2004) for example.

General Approaches to Optimal or Near-Optimal Constrained Matching

1. Calculate propensity scores
2. Establish a **distance matrix**

This is just a table with one row for each treated subject and one column for each potential control.

- The "distances" can be squared differences in propensity scores between the subjects, Mahalanobis distances, or something else.
- To use calipers, we set to ∞ all cells in the table corresponding to a propensity difference which exceeds the caliper.

A Small Distance Matrix

Consider four treated subjects (T1, T2, T3 and T4) and six control subjects (C1, C2, C3, C4, C5 and C6.)

- We have a difference score (perhaps the absolute difference in propensity for treatment) for each comparison. Some of these are infinite.
- We also have each subject categorized as (Y)oung or (O)ld, and we haven't decided yet how important this is for our matching.

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

OK, so Who Gets Matched?

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

- Now, who gets matched?

OK, so Who Gets Matched?

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

- Now, who gets matched?
- Treated subject T1 matches to C1

OK, so Who Gets Matched?

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

- Now, who gets matched?
- Treated subject T1 matches to C1
- T2 matches to C3

OK, so Who Gets Matched?

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

- Now, who gets matched?
- Treated subject T1 matches to C1
- T2 matches to C3
- T3 matches to C4 (or maybe C6 - is age important?)

OK, so Who Gets Matched?

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

- Now, who gets matched?
- Treated subject T1 matches to C1
- T2 matches to C3
- T3 matches to C4 (or maybe C6 - is age important?)
- T4 matches to C6 (or C2, or C4, hmmm....)

Almost Exact Matching

- Suppose a few of the covariates are of enormous importance - want to match exactly on them wherever possible.

We could add a penalty (but perhaps not an infinite penalty) to the distance matrix when the specified covariates fail to match, and that is the main approach that we use.

- Adding 2 to the Mahalanobis distance for mismatches roughly doubles the importance of that covariate as compared to the others, for example.

There's a lot of active work in this area developing various algorithms that permit finer control.

"Fine Balance" in Matching

- Constrain optimal matching that forces a nominal variable to be balanced, without restricting who is matched to whom.

This is especially useful if...

- you have a nominal variable with many levels
- you have a rare binary variable that is difficult to control using a distance
- you are focused on the interaction of several nominal variables

It is also possible to get specific imbalance patterns.

Fine Balance: Initial Distance Matrix

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24

Suppose we want to get optimal balance on the propensity score while matching perfectly on the age category (Y/O).

- We have 4 treated subjects (1 young, 3 old)
- We have 6 potential controls (2 young, 4 old)
- So we need to remove 1 young and 1 old in matching

Fine Balance: Augmented Distance Matrix

Subject	C1 (Y)	C2 (O)	C3 (O)	C4 (Y)	C5 (O)	C6 (O)
T1 (Y)	.23	.47	.39	∞	.51	.35
T2 (O)	.45	∞	.28	.31	.42	∞
T3 (O)	∞	.35	∞	.27	.44	.28
T4 (O)	.31	.26	.51	.29	∞	.24
Extra 1	0	∞	∞	0	∞	∞
Extra 2	∞	0	0	∞	0	0

Add 2 rows to the matrix, then run the match

- *Extra 1* pulls away one young control
- *Extra 2* pulls away one old control

The binary age category will be perfectly balanced across the matched sample, but the partners within each individual pair are not required to be in the same age category.

Fine Balance General Procedure

To get the minimum distance match with fine balance (on a nominal covariate, say GROUP)...

1. Cross tabulate GROUP with treatment indicator
2. Determine # of controls to remove from each category of GROUP to achieve perfect balance
3. Add one row for each control that must be removed, with 0 distance to its own category and infinite distance to all others
4. Find an optimal match for this square matrix
5. Discard extra rows and their matched controls

Section 6

Full Matching

Full Matching in Observational Studies

- In the past, it has been tough to implement full matching in observational studies, even though it is appealing in principle.
- Alignment of comparable treated and control subjects is as good as any alternate method, and potentially much better.
- Hansen (2004) modifies full matching with modifications to minimize variance as well as bias

In this example,

- Optimal full matching removes as much as 99% of the bias along a PS on which treated and control means are separated by 1.1 SD's.
- Reduces to insignificance biases along 27 covariates, while making use of more, not less, of the data than regression based analyses.

Hansen (2004) SAT Coaching Study

- Survey of a random sample of 1995-1996 SAT test takers about their preparation
- 12% of respondents had completed extracurricular test preparation courses
- Matching looked unattractive to the original researchers due to significant reduction in sample size, but they only considered 1:1 matching.
- Do 1:k matching options look better?

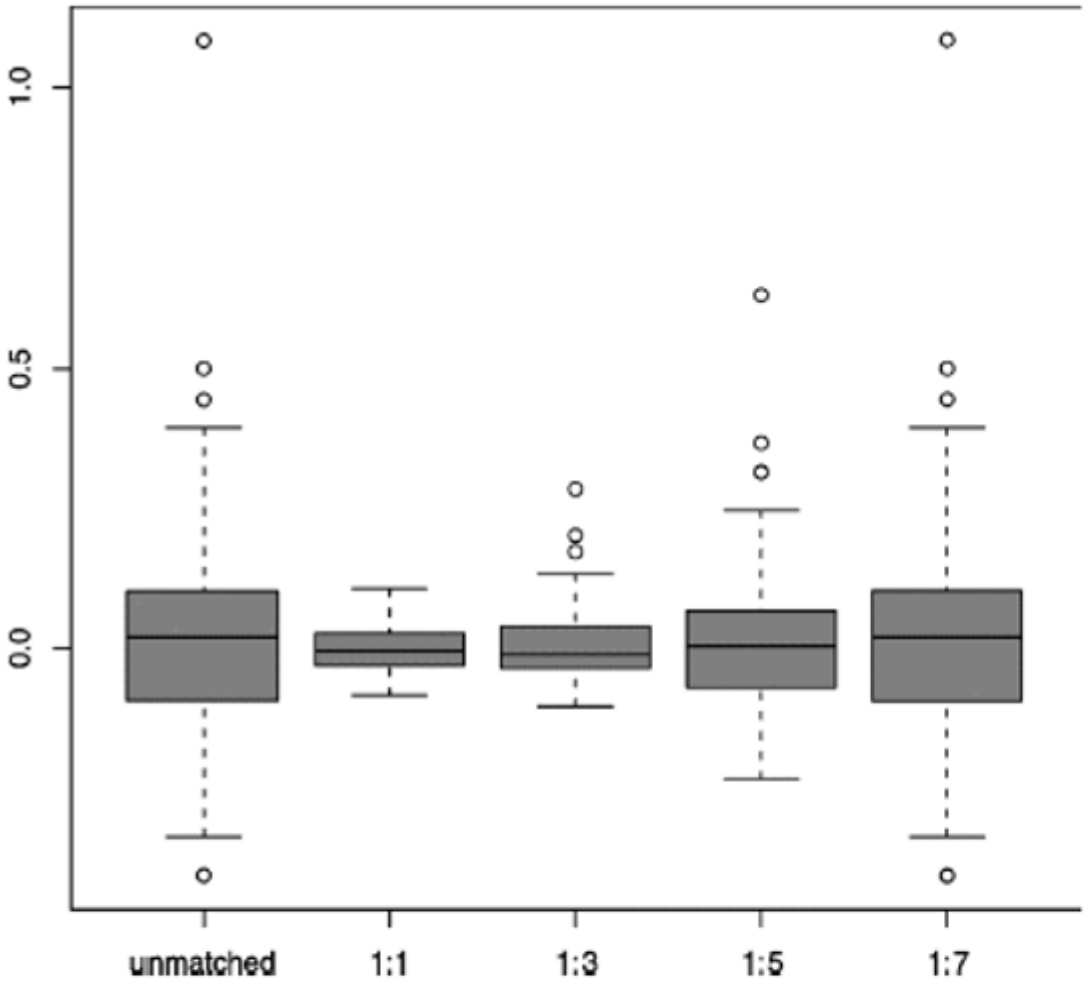


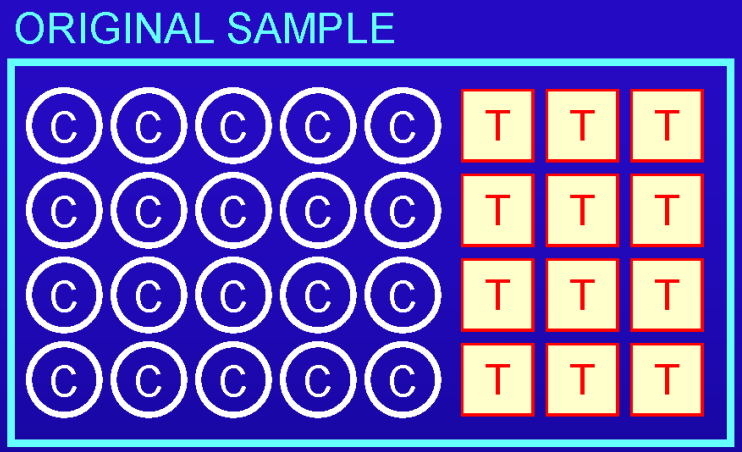
Figure 1. Covariate Imbalances in 1 : k Matching. Each boxplot represents standardized biases in the 99 categories of the 27 categorical covariates along with standardized bias in the propensity score (which in each plot is the uppermost outlier). Strictly speaking, the matching represented at far right is not a 1 : 7 matching but a blend of six 1 : 6 and 494 1 : 7 matched sets.

Covariate Imbalances in 1:k Matching

- In all of these cases, we're using less data
- Still some imbalance

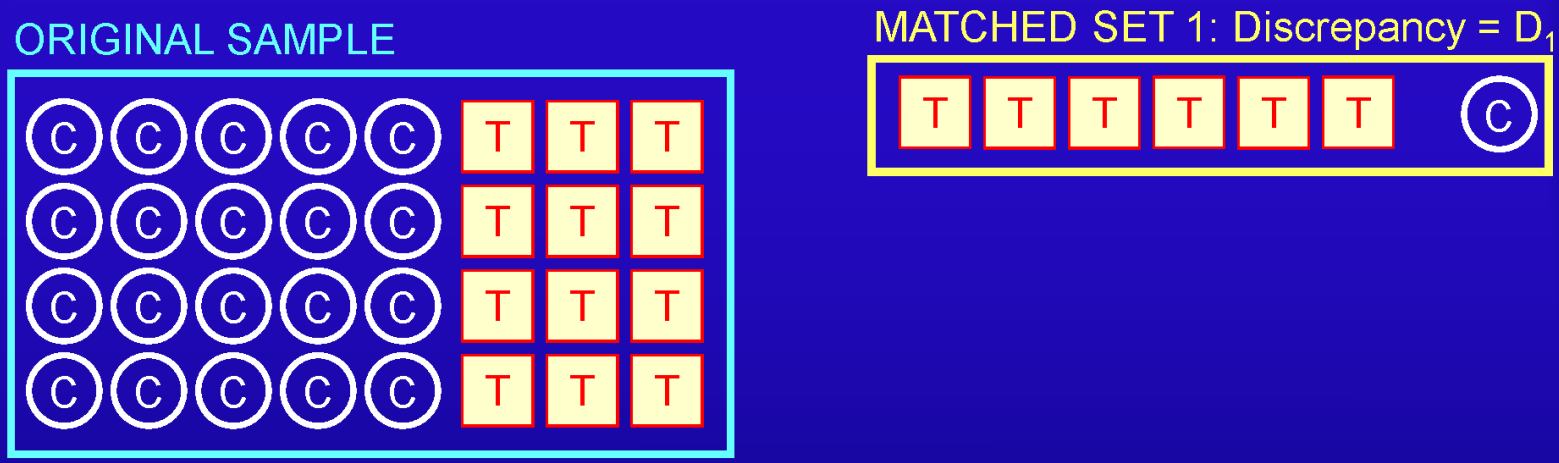
Hansen 2004

Optimal Full Matching



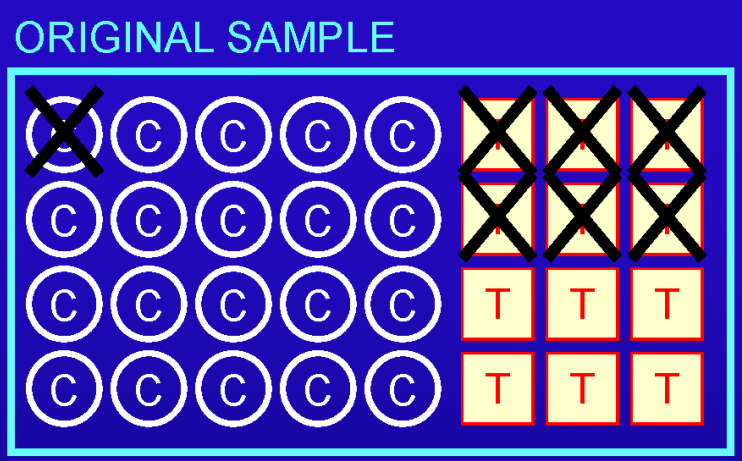
- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).

Optimal Full Matching



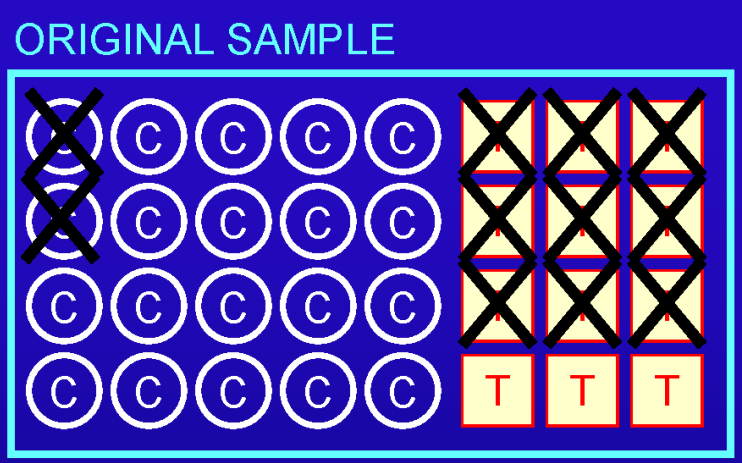
- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

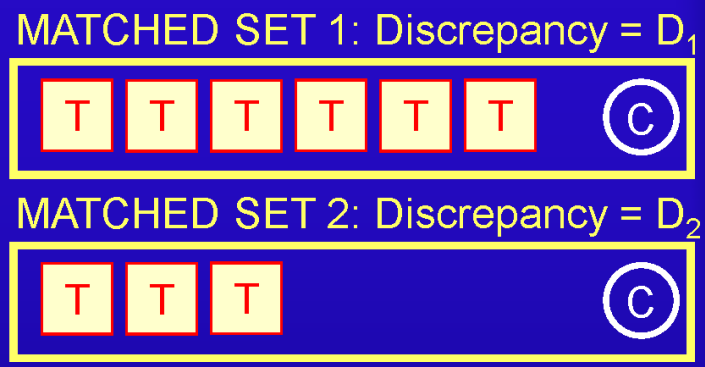
Optimal Full Matching



- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

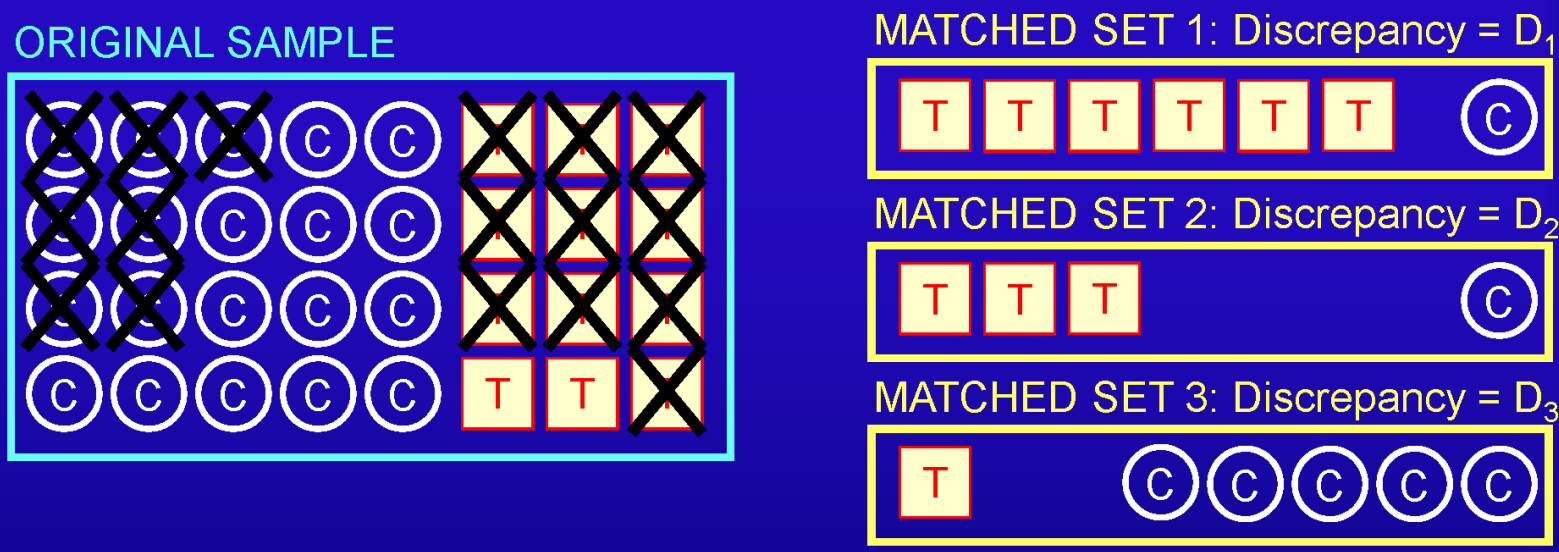
Optimal Full Matching





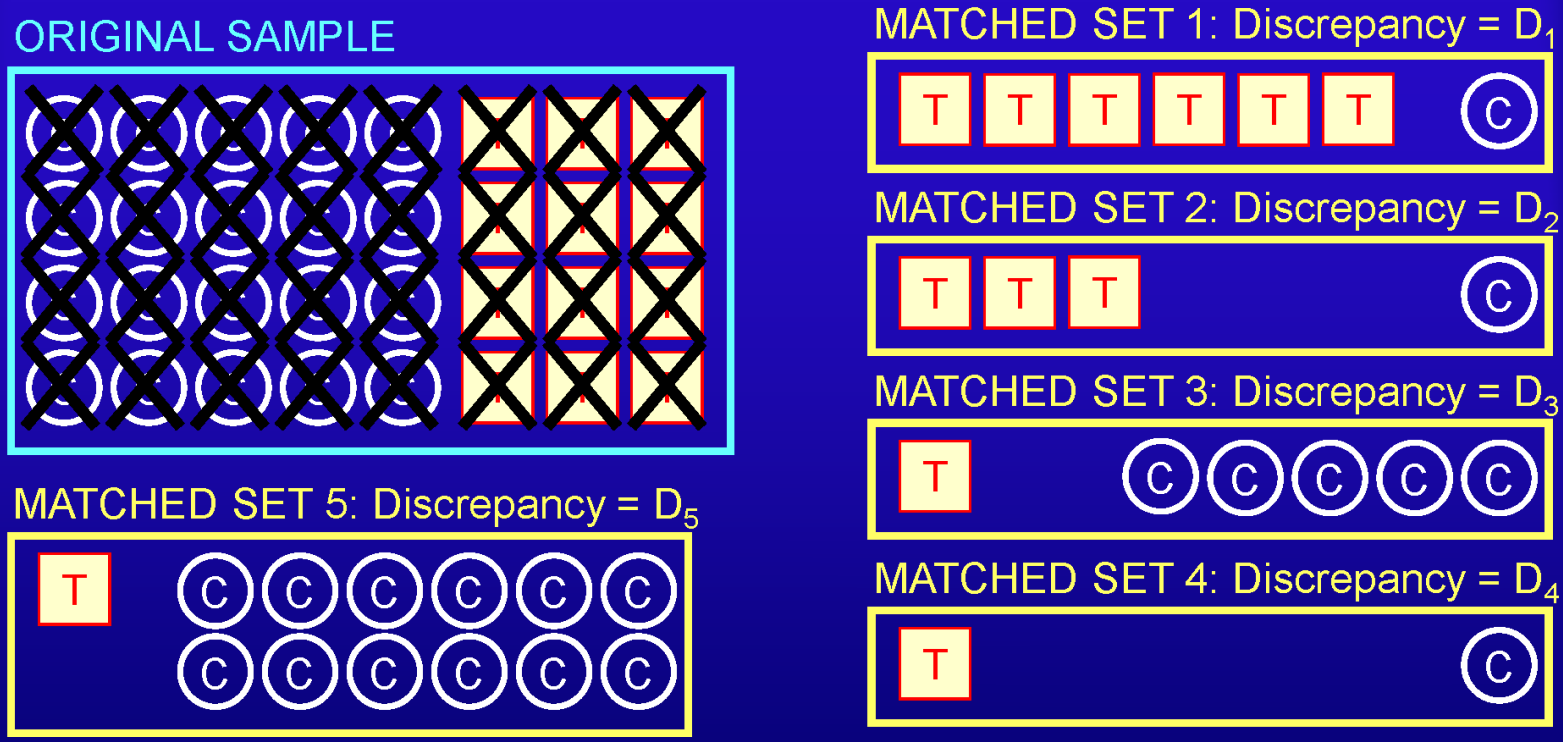
- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

Optimal Full Matching



- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

Optimal Full Matching



- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

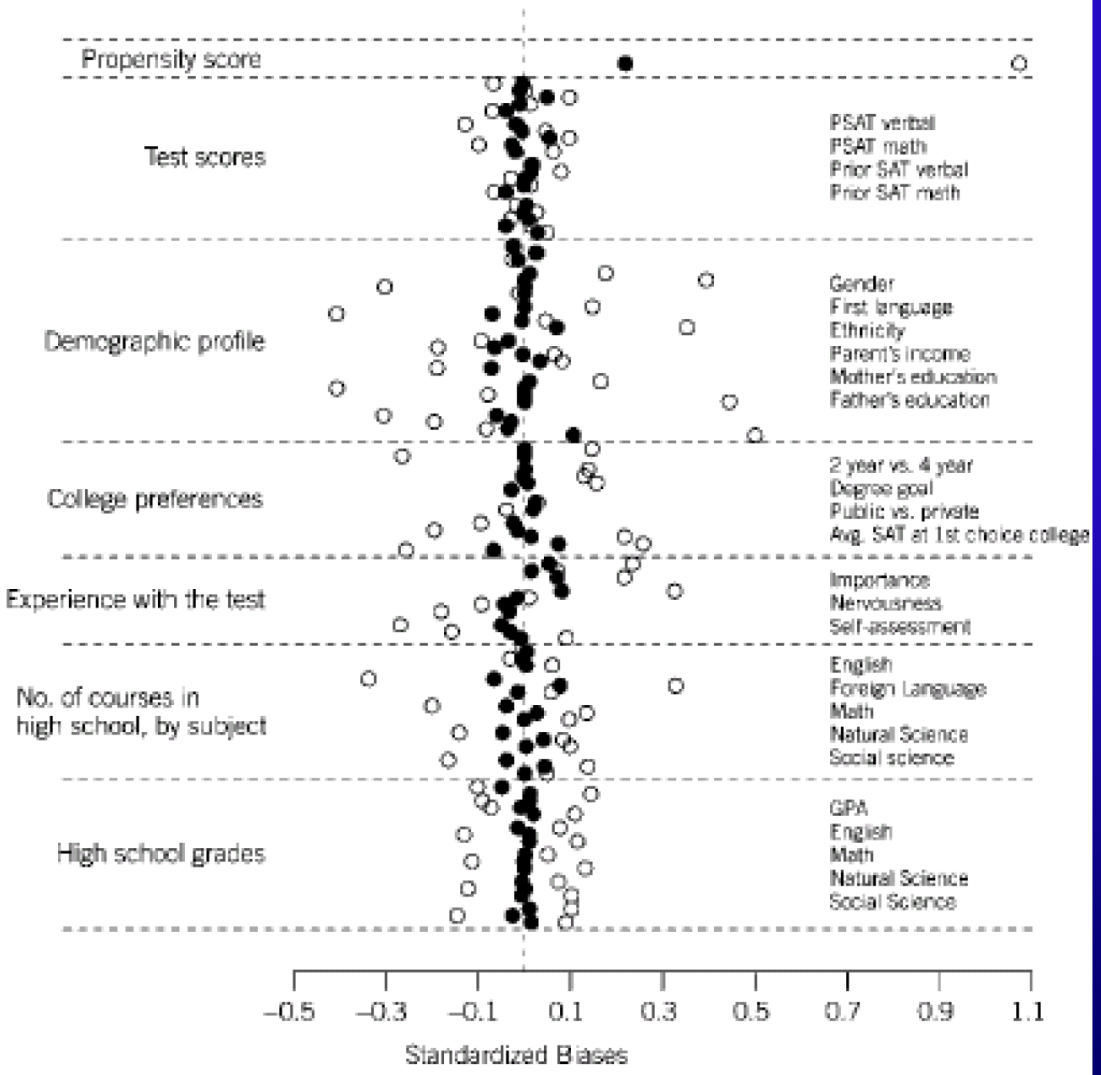


Figure 3. Standardized Biases Without Stratification or Matching, Open Circles, and Under the Optimal [.5, 2] Full Match, Shaded Circles.

Standardized Bias Plot

- Open circles are for standardized biases before matching
- Shaded circles describe results after full match

29

SAT Coaching Study Results

- Raw differences of treated and control group means were 41 points on Math and 9 on Verbal
- Full matching leads to aggregate contrasts of 26 points on Math and 1 point on the verbal.
  - Standard errors for these estimates are around 5 points.
- Surprised that Verbal effect is so small?
  - Control is not "no prep at all"
  - Estimated effect of treatment on the controls is 3 for Math and -8 on Verbal.
- Method doesn't require homogeneity of coaching effects.
- Whether and to what degree coaching is beneficial appears to vary greatly across students.

Next Time

- Designing Observational Studies (Rubin 2001)
- Discussion of Rosenbaum Chapters 5-6
- Discussion of Project Proposals
- DIscussion of Kubo et al. (2020)