500 Class 08

https://thomaselove.github.io/500-2024/

2024-03-07

What's in these Slides?

- Some Extensions to Matching (finally)
- Full Matching example
- Tanenbaum (2017) paper
- Discussion of Rosenbaum Chapter 7
- Supplement on Instrumental Variables
  - Whitehouse (2007)
  - Landrum and Ayanian (2001)
  - Posner et al. (2001)

Section 1

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 2

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?

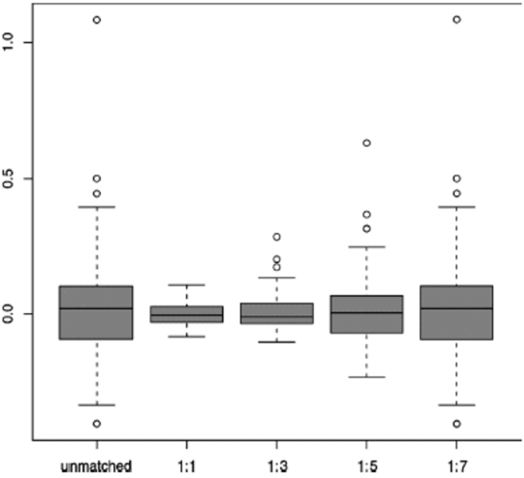


Figure 1. Covariate Imbalances in 1 : k Matching. Each boxplot represents standardized biases in the 99 categories of the 27 categorical covariates along with standardized bias in the propensity score (which in each plot is the uppermost outlier). Strictly speaking, the matching represented at far right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

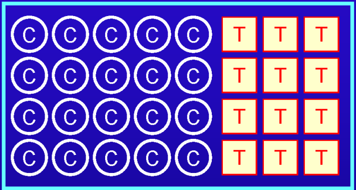
Covariate Imbalances in 1:k Matching

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

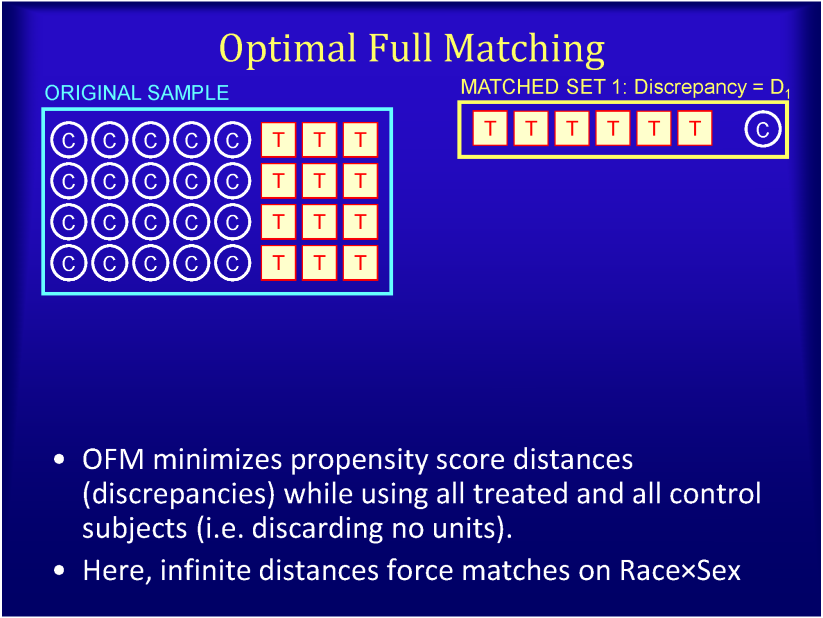
Hansen 2004

Optimal Full Matching

ORIGINAL SAMPLE



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



Optimal Full Matching

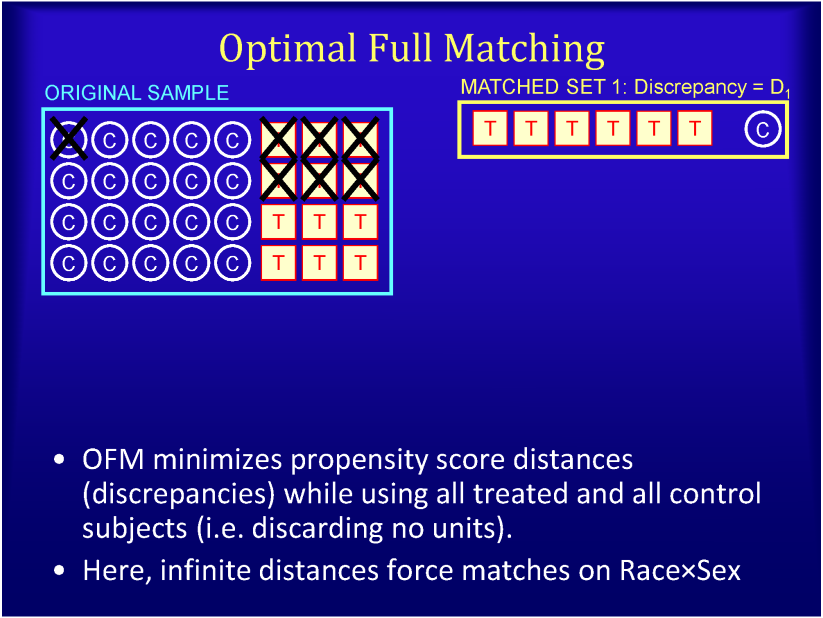
ORIGINAL SAMPLE

C C C C C T T T
C C C C C T T T
C C C C C T T T
C C C C C T T T

MATCHED SET 1: Discrepancy = D_1

T T T T T T C

- 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

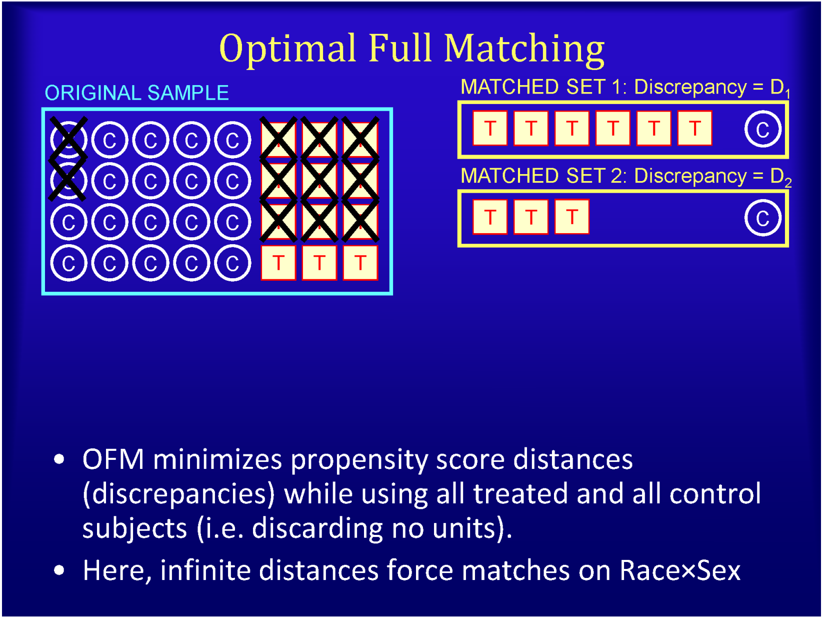


Optimal Full Matching

ORIGINAL SAMPLE

MATCHED SET 1: Discrepancy = D_1

- 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



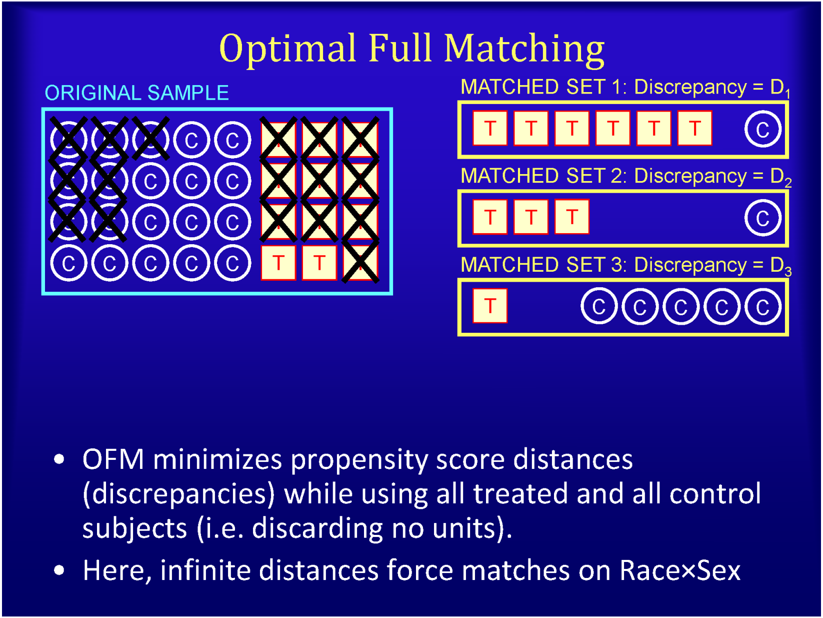
Optimal Full Matching

ORIGINAL SAMPLE

MATCHED SET 1: Discrepancy = D_1

MATCHED SET 2: Discrepancy = D_2

- 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

ORIGINAL SAMPLE

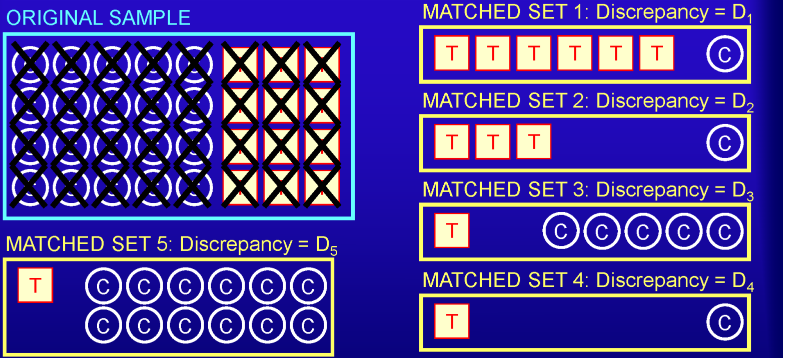
MATCHED SET 1: Discrepancy = D_1

MATCHED SET 2: Discrepancy = D_2

MATCHED SET 3: Discrepancy = D_3

- 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



ORIGINAL SAMPLE

MATCHED SET 1: Discrepancy = D_1

MATCHED SET 2: Discrepancy = D_2

MATCHED SET 3: Discrepancy = D_3

MATCHED SET 4: Discrepancy = D_4

MATCHED SET 5: Discrepancy = D_5

- 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

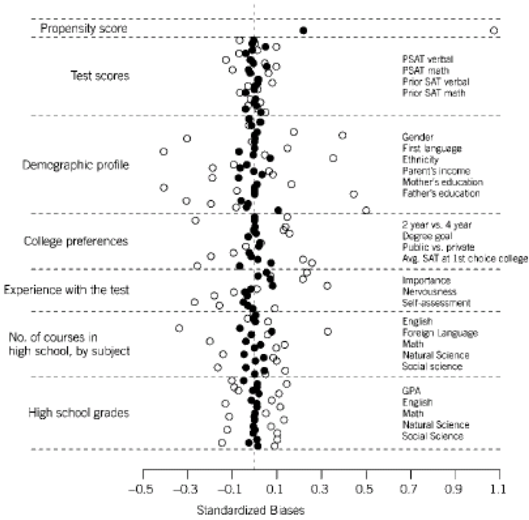


Figure 3. Standardized Biases Without Stratification or Matching, Open Circles, and Under the Optimal [.5, 2] Full Match, Shaded Circles.

Standardized Bias Plot

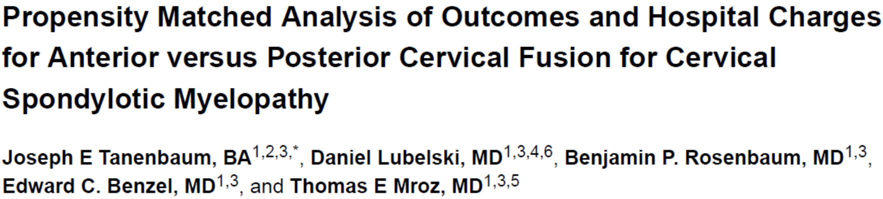
- Open circles are for standardized biases before matching
- Shaded circles describe results after full match

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.

Section 3

Tanenbaum JE et al. (2017) Propensity Matched Analysis of Outcomes and Hospital Charges for Anterior versus Posterior Cervical Fusion for Cervical Spondylotic Myelopathy



Propensity Matched Analysis of Outcomes and Hospital Charges for Anterior versus Posterior Cervical Fusion for Cervical Spondylotic Myelopathy

Joseph E Tanenbaum, BA[1,2,3,*], **Daniel Lubelski, MD**[1,3,4,6], **Benjamin P. Rosenbaum, MD**[1,3], **Edward C. Benzel, MD**[1,3], and **Thomas E Mroz, MD**[1,3,5]

Abstract for Tanenbaum et al. (1/4)

STUDY DESIGN: Retrospective analysis of data from the Nationwide Inpatient Sample (NIS), a nationally representative, all-payer database of inpatient diagnoses and procedures in the United States.

OBJECTIVE: To compare anterior cervical fusion (ACF) to posterior cervical fusion (PCF) in the treatment of cervical spondylotic myelopathy (CSM).

SUMMARY of BACKGROUND DATA: Previous studies used retrospective single-institution level data to quantify outcomes for CSM patients undergoing cervical fusion. It is unclear whether ACF or PCF is superior with regards to charges or outcomes for the treatment of CSM.

Abstract for Tanenbaum et al. (2/4)

METHODS: We used NIS data to compare ACF to PCF in the management of CSM. All patients 18 years or older with a diagnosis of CSM between 1998–2011 were included.

ACF patients were matched to PCF patients using propensity scores based on patient characteristics (number of levels fused, spine alignment, comorbidities), hospital characteristics, and patient demographics.

Multivariable regression was used to measure the effect of treatment assignment on in-hospital charges, length of hospital stay, in-hospital mortality, discharge disposition, and dysphagia diagnosis.

National Inpatient Sample (1/2)

This study used Nationwide Inpatient Sample (NIS) data from 1998–2011.

- Established by the Agency for Healthcare Research and Quality (AHRQ), the NIS is the largest all-payer healthcare database in the United States.
- The NIS contains a 20% stratified sample of all hospital discharges from 1988–2011.
- Within the database, each entry corresponds to a single hospital admission.

National Inpatient Sample (2/2)

- Using the NIS, national estimates can be generated by assigning weighted values to each hospital discharge.
- The NIS includes data on patient demographics, comorbidities, diagnoses, procedures performed, outcomes (e.g., length of hospital stay, hospital charges, mortality), and hospital features.
- Finally, the NIS codes admission diagnoses, procedures, and in-hospital complications using ICD-9-CM codes.
- In order to mitigate bias, data were used starting in 1998 because the sampling strategy of the NIS changed that year.
- Furthermore, Elixhauser comorbidity data were collected beginning in 1998.

Elixhauser Comorbidity Index

The Elixhauser Index is a composition of thirty comorbidities characterizing significant conditions associated with in-hospital mortality, including acute and chronic comorbidities.

- The Elixhauser index allows for standardized risk adjustment in administrative databases such as the NIS.
- The Elixhauser Index Score ranges from zero to thirty, with zero indicating that a patient has none of the thirty comorbidities and thirty indicating that a patient has all thirty of the comorbidities.

Matched Cohorts

We generated matched cohorts using a propensity scoring method to minimize the effect of baseline characteristic imbalances between the ACF and PCF cohorts.

- We assigned a propensity score to each hospitalization based on the likelihood of treatment using a multivariable logistic regression that included number of levels fused, spine alignment, patient demographics, hospital characteristics, and payment source as covariates.
- Similar propensity scores between the two cohorts were matched using a nearest-neighbor method within .02 standard deviations of the calculated score without replacement.

Table 1 includes about 16 covariates

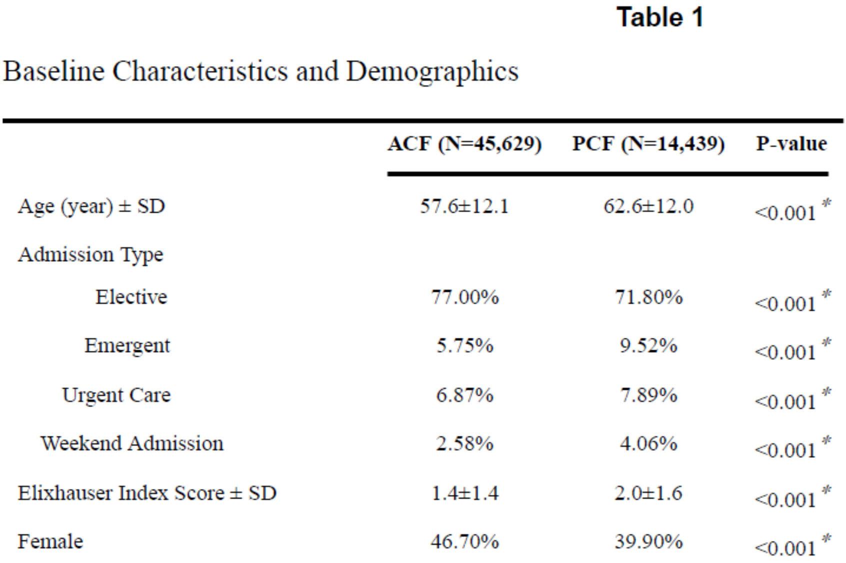


Table 1

Baseline Characteristics and Demographics

	ACF (N=45,629)	PCF (N=14,439)	P-value
Age (year) ± SD	57.6±12.1	62.6±12.0	<0.001 *
Admission Type			
Elective	77.00%	71.80%	<0.001 *
Emergent	5.75%	9.52%	<0.001 *
Urgent Care	6.87%	7.89%	<0.001 *
Weekend Admission	2.58%	4.06%	<0.001 *
Elixhauser Index Score ± SD	1.4±1.4	2.0±1.6	<0.001 *
Female	46.70%	39.90%	<0.001 *

More on Propensity Analyses

- All covariates from Table 1 (see previous slide) were used to generate a propensity score. We matched 11,671 (pairs of) PCF (and) ACF patients.
- Following the work of Rosenbaum and Rubin, we used a standardized difference with an absolute value greater than 0.10 to determine significance in assessing balance across all measured covariates.
- The variables found to be unbalanced across the two cohorts following matching were included in the final regressions.

Table 2 shows covariates after matching

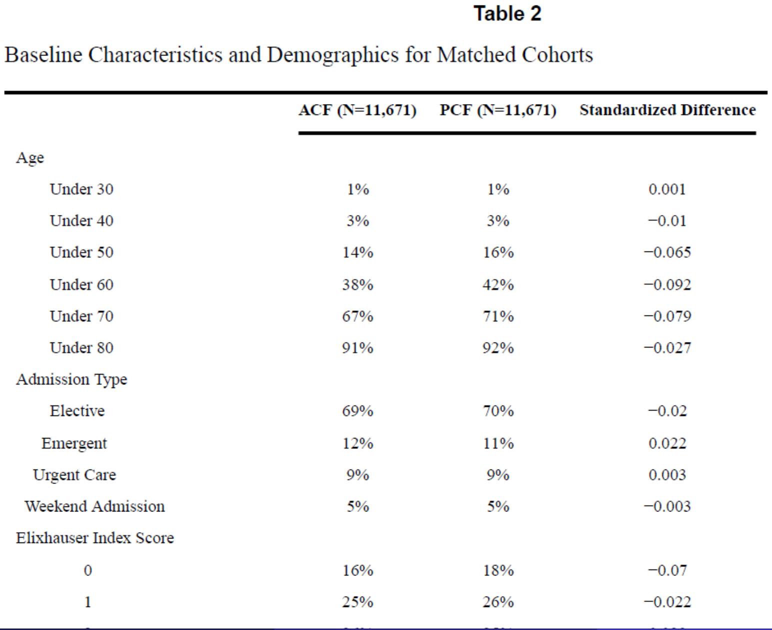


Table 2

Baseline Characteristics and Demographics for Matched Cohorts

	ACF (N=11,671)	PCF (N=11,671)	Standardized Difference
Age			
Under 30	1%	1%	0.001
Under 40	3%	3%	−0.01
Under 50	14%	16%	−0.065
Under 60	38%	42%	−0.092
Under 70	67%	71%	−0.079
Under 80	91%	92%	−0.027
Admission Type			
Elective	69%	70%	−0.02
Emergent	12%	11%	0.022
Urgent Care	9%	9%	0.003
Weekend Admission	5%	5%	−0.003
Elixhauser Index Score			
0	16%	18%	−0.07
1	25%	26%	−0.022

Five Outcome Models

- A multivariable linear regression model was used to measure the effect of treatment assignment (ACF vs. PCF) and propensity score on **LOS** and **hospital charges**.
- A multivariable logistic regression model was used to measure the effect of treatment assignment and propensity score on **in-hospital mortality**, **discharge disposition**, and **dysphagia diagnosis**.

Abstract for Tanenbaum et al. (3/4)

RESULTS: From 1998–2011, we identified 109,728 hospitalizations with a CSM diagnosis. Of these patients, 45,629 (41.6%) underwent ACF and 14,439 (13.2%) underwent PCF.

The PCF cohort incurred an average of $41,683 more in-hospital charges ($p<0.001$, inflation adjusted to 2011 dollars) and remained in hospital an average of 2.4 days longer ($p<0.001$) than the ACF cohort.

The ACF cohort was just as likely to die in the hospital (OR 0.91, 95% CI 0.68–1.2), 3.0 times more likely to be discharged to home or self-care (95% CI 2.9–3.2), and 2.5 times more likely to experience dysphagia (95% CI 2.0–3.1) than the PCF cohort.

Stability (Sensitivity) Analysis

We also performed a sensitivity analysis by first increasing and then decreasing the number of standard deviations used in the matching algorithm. These two analyses yielded similar results to those described above.

Abstract for Tanenbaum et al. (4/4)

CONCLUSIONS: In treating CSM, ACF led to lower hospital charges, shorter hospital stays, and an increased likelihood of being discharged to home relative to PCF.

Section 4

Discussion of Rosenbaum, Chapter 7

Some Rosenbaum Chapter 7 highlights

Natural Experiments, Discontinuities and Instruments

- Bits and Pieces of Random Assignment in an Otherwise Biased World
  - "Nature has its own lotteries"
- Nature's Natural Experiments
  - The Genes of Siblings (Vaidya et al. on Graves' disease)
  - Hypothetical Siblings
  - "A natural experiment is an attempt to avoid bias in treatment assignment by finding some natural setting in which treatments are nearly randomized."

Some Rosenbaum Chapter 7 highlights

- Discontinuity Designs (Thislethwaite and Campbell) as Natural Experiments
    - Assignment to treatment ends and assignment to control begins abruptly. Let's look at moments just before/after the "door slams shut."
    - DiNardo and Lee on the effect of unionization on wages
- Paul Holland's randomized encouragement experiment
    - What is the effect of encouragement? (randomized)
    - What is the effect of doing what you were encouraged to do? (trickier)
- Instrumental Variables and the Complier Average Causal Effect
    - Angrist, Imbens and Rubin set out five main characteristics/assumptions of a solid instrument
    - "We cannot recognize a complier when we see one."

Rosenbaum, Chapter 7

1. What was the most important thing you learned from reading Chapter 7?
2. What was the muddiest, least clear thing that arose in your reading?

Section 5

Supplement: On Instrumental Variables

What's in this supplement?

- Instrumental Variables
  - Is An Economist Qualified to Solve Puzzle of Autism? (Whitehouse 2007)
  - The Exclusion Restriction
  - Propensity Scores and Instruments, Together
  - The Formalized Instrumental Variables Assumptions
- Landrum and Ayanian Comparing methods (2021)
- Posner et al comparing methods (2001)

Is An Economist Qualified to Solve Puzzle of Autism?

By Mark Whitehouse, Wall Street Journal, page A1, Feb 27 2007

> *In the spring of 2005, Cornell University economist Michael Waldman noticed a strange correlation in Washington, Oregon and California. The more it rained or snowed, the more likely children were to be diagnosed with autism.*

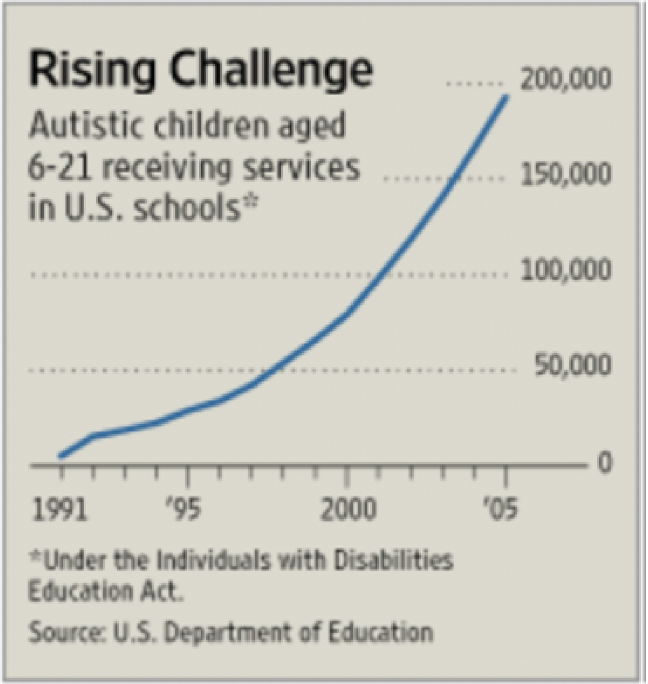
- What do kids do more during rain or snow that influences autism?

Is An Economist Qualified to Solve Puzzle of Autism?

By Mark Whitehouse, Wall Street Journal, page A1, Feb 27 2007

> *In the spring of 2005, Cornell University economist Michael Waldman noticed a strange correlation in Washington, Oregon and California. The more it rained or snowed, the more likely children were to be diagnosed with autism.*

- What do kids do more during rain or snow that influences autism?
- Watch TV?



Rising Challenge

Autistic children aged 6-21 receiving services in U.S. schools*

*Under the Individuals with Disabilities Education Act.

Source: U.S. Department of Education

Autism on the Rise

Studies in recent decades have shown the proportion of children with autism growing, though researchers aren't sure the disorder has actually become more prevalent. Greater awareness, broadening definitions of the disorder and the availability of special-education programs may have made parents more likely to get their children diagnosed.

Most researchers now recognize that heredity plays a central role in autism, and they are making progress in identifying the genes responsible. They're also looking into the possibility of interaction with environmental factors, both in the womb and after birth.

Professor Waldman's Interest in Autism

Professor Waldman's 2-year-old son was identified with an autism-spectrum disorder in 2003.

Hoping to eliminate potential triggers, Professor Waldman supplemented recommended therapy with a sharp reduction in TV watching. His son had started watching more TV the previous summer, after a baby sister was born.

Waldman's son improved within six months and today has fully recovered – "When I saw the rapid progress, which was certainly not what anyone had been predicting, I became very curious as to whether television watching might have played a role in the onset of the disorder."

Does TV Trigger Autism?

- Ideal Study: randomly select a group of susceptible babies at birth to refrain from TV
    - Compare their autism rates to a control group
- Economists look for “natural experiments” and use **instrumental variable** methods to hopefully approximate the rigor of a randomized trial.
- Think of an instrument as a randomized “nudge” towards a treatment...

Levitt and Dubner, *Freakonomics* and *SuperFreakonomics*, for instance, provide popular treatments of some key ideas.

What is an Instrumental Variable?

- An instrumental variable is a “random”, “policy” or “natural” nudge or encouragement towards a particular exposure that affects the outcome only to the extent that it affects acceptance of the exposure.
- An ideal instrumental variable to test the link between an exposure and an outcome has...
  - a strong correlation with receipt of the exposure
  - no direct effect on the outcome or on other factors that cause the outcome (exclusion restriction)

Then, if data links the IV to the outcome, it suggests that the *exposure* must be contributing to the outcome.

Back to Autism for a Moment (from Whitehouse, WSJ)

> *In principle, the best way to figure out whether television triggers autism would be to do what medical researchers do: randomly select a group of susceptible babies at birth to refrain from television, then compare their autism rate to a similar control group that watched normal amounts of TV. If the abstaining group proved less likely to develop autism, that would point to TV as a culprit.*
>
> *Economists usually have neither the money nor the access to children needed to perform that kind of experiment… Instead, economists look for instruments – natural forces or government policies that do the random selection for them.*

On Instruments (from Whitehouse, WSJ)

> *First developed in the 1920s, the technique helps them separate cause and effect.*
>
> *Establishing whether A causes B can be difficult, because often it could go either way. If television watching were shown to be unusually prevalent among autistic children, it could mean either that television makes them autistic or that something about being autistic makes them more interested in TV.*
>
> *The ideal instrument is a variable that is correlated with A but has no direct effect of its own on B. It should also have no connection to other factors that might cause B. If data in a study nonetheless show that the instrumental variable is linked to B, it suggests that A must be contributing to B.*

A Simple Example

This example largely comes from https://www.statology.org/instrumental-variables/.

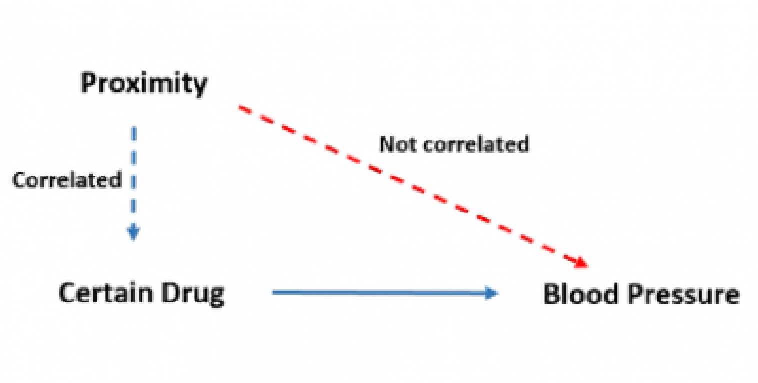
1. Suppose we want to estimate the effect that a certain drug has on systolic blood pressure.
2. Other variables like time spent exercising, overall diet, and stress levels also affect systolic blood pressure.
3. So if we run a simple linear regression to predict SBP based on using the drug or not using the drug alone, we can't be sure that we have accurately captured the effect of the drug on SBP.

Identifying an Instrument

An instrumental variable is a third variable introduced into regression analysis that is correlated with the predictor variable, but uncorrelated with the outcome variable. Our goal is to estimate the true causal effect that our exposure has on our outcome.

An example of an instrumental variable that we may use in this regression analysis is an individual's proximity to a pharmacy.

- "proximity" would likely be highly correlated with whether or not the individual takes the certain drug because an individual wouldn't be able to obtain it if they don't live near a pharmacy.
- But "proximity" isn't expected to have any correlation with SBP (our outcome) except through the exposure (taking the drug.)


Proximity
Correlated
Not correlated
Certain Drug
Blood Pressure

Instrumental Variables Regression (Two-Stage Least Squares)

1. Fit a regression using the instrument as the predictor of our exposure.

Certain Drug $= \beta_{01} + \beta_{11}$ (Proximity)

Predicted Values are called $\hat{cd}$.

Stage 2 of the IV Regression

2. Fit a regression using the predicted values of our exposure from model 1 as the predictor for our outcome.

SBP $= \beta_{02} + \beta_{12}\ \hat{cd}$

Because we used “proximity” alone to come up with $\hat{cd}$ and since we know that proximity should not be correlated with SBP, the effect seen in the second stage regression can be attributed to the drug.

So the effect captured in β_{12} then tells us about a causal effect of the drug on SBP.

Key Assumptions in this Simple Example

1. Our instrument (proximity) is highly correlated with our exposure (use of the drug.)
2. Our instrument (proximity) is not correlated with our outcome (systolic blood pressure.)
3. Our instrument (proximity) is also not correlated with the other variables that are left out of the model (like exercise, diet, stress.)

If the instrument does not meet these criteria, it will likely produce unreliable and biased results.

Instruments have changed the world

> *Joshua Angrist (and others) have studied the cost of war. Steven Levitt has examined the effect of adding police on crime (for example.) Their work has played an important role in public-policy debates.*

Angrist's work (with Imbens and Rubin in 1990) is perhaps the most cited.

> *Did service during the Vietnam War have a negative effect on people's future earnings?*
>
> *It wouldn't be enough to say that people who served ended up poorer. Perhaps a lack of opportunities in the civilian world made them more likely to enlist in the first place.*

Angrist (1990)

> *As an instrumental variable, Professor Angrist chose the draft lottery, which made some people more likely than others to serve in the Vietnam-era military, but didn't have any connection to their initial circumstances.*

- Data: On average, white men with draft-eligible lottery numbers had much lower earnings many years later.
- Data on non-whites were inconclusive.

  Prof. Angrist concluded that conscription had a detrimental effect on future earnings.

Also see Angrist, Imbens and Rubin (1996, JASA)

Levitt (1997)

Chicago's Professor Levitt tackled police staffing and crime. That's an issue where cause and effect are hard to disentangle because cities with many criminals are likely to have more police, but that doesn't mean an excess of officers causes crime.

Prof. Levitt took advantage of the fact that mayors and governors tend to put more police on the streets in election years. Using election cycles, he concluded in a 1997 paper that adding police reduces violent crime.

Understanding the Exclusion Restriction for an Instrument

Randomized encouragement to either an active drug or a double-blind placebo is the experimental design that most closely approximates an instrument.

- Encouragement is actually randomized.
- Neither subject nor investigator knows what treatment subject is being encouraged to do.
- So there are few opportunities for encouragement to affect a clinical outcome without shifting the amount of active drug that is consumed.

See Rosenbaum (2010) or Greevy et al (2004) or Holland (1998)

Instrumental Variable for Looking at TV-Autism relationship

Waldman et al. selected precipitation

- Kids tended to spend more time in front of TV when it rained or snowed than when it didn't.
- IV argument: Precipitation "randomly selects" some kids to watch more TV than others.
- Study conducted in WA, OR and CA where rain and snowfall vary a lot.
- Kids growing up in periods of unusually high precipitation were in fact more likely to be diagnosed with autism.

A second instrument

- Communities with larger rates of households subscribing to Cable also had higher autism rates.

Conclusions

- TV watching could be a cause of autism.
- Precipitation could be linked to potential triggers other than TV watching (household mold?)
- Marginal Effect: data reflect TV effect on the kids who changed their habits because of rain or snow.
- Does nothing to explain the mechanism by which TV would influence autism, as in all IV studies.

The Instrumental Variable Idea

Find a variable (the instrument)

- strongly correlated with the treatment choice
- but having no direct effect on the outcome (outside of the instrument's influence on treatment selection)

If these two conditions are not met, then IV is not a useful approach.

- In health care, treatment selection is usually closely linked to outcome.

See Earle et al (2001), Landrum and Ayanian (2001), Posner et al (2001)

Section 6

Landrum Mary Beth and Ayanian John Z 2001 Causal Effect of Ambulatory Specialty Care on Mortality Following Myocardial Infarction: A Comparison of Propensity Score and Instrumental Variable Analyses

Landrum and Ayanian (2001)

Propensity Scores and Instrumental Variables Together: Ambulatory Specialty Care following an Acute MI

- Landrum and Ayanian (2001) studied the effect of specialist (cardiologist) vs. generalist outpatient care for acute MI pts.
- Patients getting specialty care for AMI were younger, less likely to have chronic illnesses, and more likely to have prior cardiac disease.
  - Data from Cooperative Cardiovascular Project
  - 200,000 Medicare AMI patients treated in 1994-5.
  - Especially detailed clinical data available in several states: focus here is on NY fee-for-service patients.

Landrum and Ayanian (2001) Design

- Outcome: 18m mortality after AMI discharge
- Treatment: Cardiology vs. Generalist Care
    - 3,551 (65%) cardiology care: had at least one office visit with cardiologist in 90d post-discharge
    - 1,916 (35%) generalist care: had at least one PCP office visit without any cardiologist visits in 90d

Unadjusted mortality was substantially lower (9.3% vs. 15.8% at 18m) in cardiology group

- But the two groups were very different in terms of important baseline covariates linked to mortality…

Landrum and Ayanian (2001) Cardiology patients were:

- Younger (mean 73.4 vs 74.5, stdzd diff = 20%)
- More often male, white and to have a prior history of AMI or angina than generalist patients
- Less likely to also have stroke, COPD or diabetes or prior CHF than generalist patients
- More likely to be discharged from:
    - A teaching hospital and also an urban hospital
    - A hospital with invasive cardiac services
    - A hospital with cardiology care facilities, or angiography, or angioplasty, or bypass surgery facilities

Standard Logistic Regression Analysis

- Adjusting for these (and more) observed differences in a logistic regression model reduced the unadjusted absolute differences in 18-month mortality from 6.5% to 2.9%.
- But given the substantial differences in observed characteristics, and the likelihood that patients differ in terms of unobserved characteristics related to outcome as well…

Propensity Scores for the Ambulatory Care Study

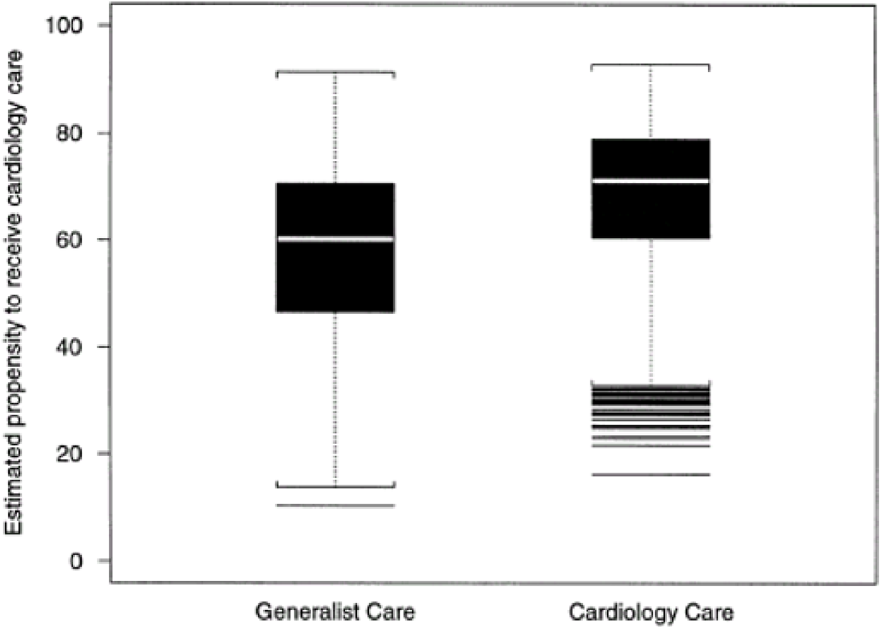
- Logistic regression predicting receipt of cardiology care in 90 days following discharge.
- 41 clinical and provider characteristics included…
  - All those covariates described previously
  - Patient demographic and clinical characteristics
  - Treatments received during hospitalization
  - Inpatient provider and hospital characteristics

The propensity model did not include the "instrument" we will discuss shortly.

Overlap in Propensity Score shown in next slide:

- 34 patients got cardiology care despite a low PS.
- Did we not observe all predictors of cardiology care?

Overlap in Propensity for Cardiology Care



Step 1. Propensity Score Stratification

- Balance achieved on most covariates: exceptions are % rural, history of HTN, in-hosp. cardiac arrest
- In patients least likely to receive cardiology care (by PS quintile) cardiology care estimated to reduce 18m mortality by 11.5 points (12.3% vs 23.8%)
    - Other quintiles show much smaller differences
- Average of differences = 3.1 points (est. mortality reduction if all pts. in cohort got cardiology care)
- Differences weighted by cardiology pts in quintile = 2.3 points (average causal effect among the treated)
- Results insensitive to using 3 or 10 strata instead of 5

Step 2. Propensity Score Matching

- 1775 of 1916 generalist patients (93%) matched to a cardiology patient with closest estimated PS [inside 0.6 SD of logit PS caliper]
- 1776 unmatched cardiology patients were those with highest propensity for cardiology care
- Covariate Balance is excellent among matches

Among matched patients,

- 18 m mortality for cardiology patients was 11.7%
- 18 m mortality for generalist patients was 14.7%
- 3.0 point absolute reduction in mortality for cardiology care (standard error = 1.1)

Step 3. Instrumental Variable

Tough part: identifying the instrument

- Related to treatment but not outcomes
- Selection: density of cardiologists in patient's county of residence, at two levels (above or below 6.7 cardiologists per 100,000 population age 65+)

We'll obtain an estimate of the Local Average Treatment Effect for all "Marginal" Patients (also called "Compliers")

- Would get cardiology if they lived in a high density area but not if they lived in a low density area
- Cannot identify "compliers" from observed data

Formalized Instrumental Variables Assumptions

1. SUTVA (Stable Unit Treatment Value Assumption) - unaffected by other subjects
2. Non-zero causal effect of instrument on treatment (IV predicts treatment status)
3. Ignorable assignment of the instrument
4. Exclusion Restriction (IV has no effect on outcomes other than through the treatment)
5. Monotonicity of IV's effect on treatment

The SUTVA Assumption

- A patient’s potential treatments / outcomes are assumed unrelated to treatment status of other patients
  - Treatment Status (cardiology / generalist) and Mortality assumed unaffected by the care received by other patients
  - Access to care does vary across geographical areas
  - Patients in high-density areas may have increased access to all kinds of specialists

Non-Zero Causal Effect of Instrument on Treatment

- The IV must predict treatment status
  - We can check this (to some degree) in data.
  - Likelihood of receiving cardiology care was positively associated with quintile of cardiology density ($p < 0.001$) in New York.
  - This wasn't true in other states (TX, CA, MA) so this instrument wouldn't be appropriate there.

Ignorable Assignment of the Instrument

- Patients from different density areas must be similar (in both observed and unobserved characteristics) to what they would have been had density been randomly assigned.
  - Can't verify directly, but if patients are similar in terms of observed characteristics, that provides some evidence of the validity of the assumption.
  - In fact, observed data in NY looks balanced when we compare low density to high density patients.
  - Not true in other states (FL high density areas had older patients, for instance)

Exclusion Restriction

- IV assumed to have no effect on outcomes other than through its effect on the treatment
- Can't verify this directly, either, but ...
  - Density is pretty strongly correlated with hospital characteristics and with in hospital treatments.
  - Instrument could have an effect on mortality through these other characteristics / treatments.
  - High density areas were more urban, and care processes differ between urban and rural areas.

Monotonicity

- IV assumed to affect treatment monotonically
- Can't verify this directly, either.
    - If a patient in a low-density area received cardiology care, have to believe (s)he would also have received cardiology care if (s)he lived in a high density area.
- Often seems pretty reasonable with this sort of instrument.

Analytic Decisions in Landrum and Ayanian

- Divided patients into quintiles according to density of cardiologists in their county
- Non-parametric approaches to estimating treatment effects: avoid further assumptions
    - Density strongly related to urban/rural location, so also estimated these effects separately
    - Hospital (teaching or not) and inpatient treatment (coronary angiography or not) also stronly correlated with density, so estimated treatment effects at fixed levels of those characteristics as well.

Landrum and Ayanian IV Results

LATE estimate = difference in 18m mortality (cardiology - generalist) among patients for whom cardiologist supply determined treatment.

- IV Model with No Covariates: LATE = -9.5%, SE = 7.9
- IV Model with Covariates: LATE = -1.0%, SE = 8.4

Covariates controlled for teaching hospital and inpatient treatments.

- SEs are large because treatment determined by supply for only about 15% of population.
  - Lowest Quintile of Density: 57% got cardiology
  - Highest Quintile of Density: 72% got cardiology

Comparison of PS and IV Approaches

- PS analyses found a small but significant benefit, concentrated among patients with lowest propensity to receive cardiology care.
- IV point estimates were consistent with a small benefit to ambulatory cardiology care, but were not precisely estimated, so the differences between groups were not statistically significant.

Issues to Consider

- Both PS and IV approaches rely on critical and untestable assumptions.
- They are looking at different things: if there is heterogeneity in the impact of cardiology care across strata of patients, PS and IV estimates of causal effects may differ even if both sets of assumptions hold up.
- Methods estimate effects for different people.

Importance of the Policy Question

- PS analysis lets us identify characteristics of the population to make recommendations for individual patients (subgroup analyses, etc.)
- IV analysis more applicable if we want to look at, say, the impact of increasing the supply of cardiologists, because IV demonstrates the marginal effect of such changes.

When are IV methods especially attractive?

1. An instrument is available, and...
2. Assignment to a treatment is ignorable, but compliance with the assignment is not perfect so that the dose of treatment received is non-ignorable.
3. Data are weak, in the sense that observed covariates provide insufficient insight into the background to allow estimated effects (adjusting for covariates) to be due to treatment.

- An interesting perspective on the roles of IV and PS, (and a nice application of PS with subclassification) is provided in Coyte et al.'s joint replacement study (2000) J of Health Econ.

Propensity Scores vs. Instrumental Variables?

- Both propensity score methods and instrumental variables (IV) can be used to adjust for unobserved covariates that affect treatment assignment when treatment assignment and treatment outcome are confounded.
- Some questions call for PS adjustment, others for IV models of Rx effect.
- Each have unverifiable assumptions:
  - PS adjusts for selection bias in terms of identified covariates - we must presume this is sufficient to also adjust for unobserved covariates. Sensitivity analysis can help.
  - IV presumes we can and do identify appropriate instrument(s).

Section 7

Posner MA Ash AS Freund KM Moskowitz MA Shwartz M 2001 Comparing Standard Regression, Propensity Score Matching and Instrumental Variables Methods for Determining the Influence of Mammography on Stage of Diagnosis. Health Services and Outcomes Research Methodology

Goals of Posner et al. 2001

- Mammography screening and its effectiveness in detecting cancer at an earlier stage.

Compare results of three analytic approaches:

1. Standard (regression-based) adjustment for baseline risk plus a treatment indicator
2. Propensity score matching to account for selection bias through evening out covariate distributions
3. Instrumental variables to address unmeasured differences between treated and untreated patients

The Research Question

Use of mammography for screening women over age 70; as of 2001, the value hasn't been established

- Most RCTs of mammography include no women over age 70 (focus is on the 50-70 year olds)
- No RCT has reported age-specific data within the 50-70 age group so that trends can be studied
- Breast cancer incidence continues to rise beyond age 65 - 48% of new cases are > 65.

The Data

Linked Medicare - Tumor Registry Database

- Sample consisted of all women with a first diagnosis of breast cancer …
    - In one of three regions (metropolitan Atlanta, state of Connecticut, or Seattle-Puget Sound)
    - whose utilization of mammography could be tracked for the 2 years prior to the diagnosis of breast cancer
    - who were either regular mammography users or mammography non-users (excluded "tweeners")

Treatment Variable

- Regular mammography users had claims for two separate bilateral mammograms within the two years prior to their breast cancer diagnosis, which were at least 10 months apart.
- Non-users were women with no mammography claims in the two years prior to their diagnosis.

Primary Outcome

Stage at diagnosis, dichotomized

- Early (in situ, or Stage I)
- Late (Stage II, III or IV)

Excluded the 7.4% of women with unstaged cancer

Covariates

- Age at diagnosis
  - Categorical: 67-69, 70-74, 75-79, 80-84, 85+
- Comorbidity (Charlson Comorbidity Index)
- Race (black vs. other)
- Income (median income of patient's zip code)
  - Dichotomized to highest 40% vs. lowest 60% of incomes within each region
- # of claims for primary-care office visits over the last two years (also categorized)

Approach 1: Risk Adjustment

Developed a logistic regression model to predict stage at diagnosis (early or late) from user status (regular user or non-user), controlling for:

- Region, Age, Race, Comorbidity, Median income [zip code], Primary care visits

Conclusion

Regular users have **2.97** times the odds of being diagnosed at an early stage relative to non-users (95% CI: 2.56, 3.45)

Approach 2: Propensity "Matching" (sort of)

Propensity model included same covariates as risk adjustment model:

- Region, Age, Race, Comorbidity, Median income [zip code], Primary care visits

Steps:

1. Split data into deciles based on propensity score
2. Within each decile, take a random sample from the larger group (users or non-users) to get the same number as in the smaller group
3. Matched sub-samples combined to yield final data set

I'd call this "Stratification" more than "Matching"

Propensity “Matching” inside Deciles

Decile	Non-Users	Users	*Matched* Non-Users	*Matched* Users
1	416	57	57	57
2	339	89	89	89
3	359	136	136	136
4	239	205	205	205
5	193	289	193	193
6	159	277	159	159
7	145	347	145	145
8	96	327	96	96
9	113	394	113	113
10	81	395	81	81

Covariate Balance Pre- and Post-“Matching”

Variable	Pre-match p	Post-match p
Age at diagnosis	0.001	0.98
Comorbidity Index	0.001	0.73
Race	0.001	0.35
Income	0.061	0.49
Primary Care Visits	0.001	0.51
Location (Region)	0.001	0.98

- And, looking at our outcome …

Variable	Pre-match p	Post-match p
Stage of Cancer	0.001	0.001

Results from Propensity Analysis

- Most extreme propensity scores were examined, and were close to the others, so no pairs were excluded on that basis.
- Balance dramatically improved (in terms of significance) for all variables.

Conclusion

Regular users have **3.24** times the odds of being diagnosed at an early stage relative to non-users.

- 95% CI for odds ratio: (2.69, 3.88)

> *[The propensity] approach estimates the impact of being a user of mammography for the population whose measured covariates conform to the matched sample … This result being so close to that of the standard model provides some reassurance that the standard model has adjusted correctly for any differences in measured covariates between the user and non-user groups.*

Approach 3: Instrumental Variables

Which variable to use as the instrument? We need:

1. An association between the instrument and the exposure (must predict user status)
2. **AND** a lack of correlation between the instrument and the unmeasured covariates that are associated with the outcome.
    - no residual predictive power on stage at diagnosis, after controlling for the other covariates in the model

Region as the Instrument

Trichotomous variable (Atlanta, Seattle, Connecticut)

1. Is there an association between region and use of mammography?
    - Literature suggests that there is.
    - These data seem to back the claim up.

Region as Instrument?

2. Is there no correlation between region and the unobserved covariates associated with the outcome (once we've adjusted for observed covariates in the model)?
   - Cannot test this statistically.
   - "Seems reasonable" that outcome for someone using mammography in one region shouldn't differ from outcome for someone of similar characteristics using mammography in another.

The Detailed Argument

- We have to agree that we would expect that a woman with certain characteristics (age, race, etc.) receiving regular screening in Seattle would have the same likelihood of early stage disease diagnosed from mammography had she lived in Atlanta or Connecticut.
- If this is not true, implies that follow-up after a positive mammogram differs by region.

Two-Stage Model for Instrumental Variables Approach

1. Predict user status using covariates and the instrument(s).
    - Obtain predicted probability of mammography use for each subject.
2. Predict stage at diagnosis (early or late) using the usual covariates (not including the instrument) along with the predicted probability of mammography use (instead of actual user status).

Instrumental Variable Results

- Precision will be drastically reduced from what we've seen in the previous analyses.
    - Replacing 0/1 user status with a prediction that can vary across (0, 1).

Conclusion

Regular users have **3.01** times the odds of being diagnosed at an early stage relative to non-users.

- 95% CI for odds ratio: (1.09, 8.34)

Comparison of Approaches

We start with the **standard analysis**, a logistic regression predicting stage at diagnosis that includes as independent variables a set of covariates to adjust for differences in baseline risk plus an indicator variable for whether the woman used screening. Next, we employ **propensity score matching**, which evens out the distribution of measured baseline characteristics across groups, and is more robust to model misspecification than the standard analysis. Lastly, we conduct an **instrumental variable** analysis, which addresses unmeasured differences between the users and non-users.

Approach	*OR*	95% CI
Risk Adjustment	2.97	2.56, 3.45
Propensity "Matching"	3.24	2.69, 3.88
Instrumental Variable	3.01	1.09, 8.34

$OR =$ odds of regular users being diagnosed at an early stage relative to non-users

Posner et al. Conclusions (1/2)

In summary, all three analyses - the standard regression, the propensity score matching, and the instrumental variable analysis using region as the instrument - produced very similar results. The similarity of these results helps strengthen the credibility of the standard regression analysis. There is little model mis-specification, either from measured variables, as seen via the propensity score matching, nor from unmeasured variables (that meet the instrumental variable criteria), as seen via the instrumental variable analysis.

Posner et al. Conclusions (2/2)

We recommend that investigators analyzing administrative databases or other observational studies consider the sources of bias that may affect their results. … It is important to look beyond the standard analysis and to consider propensity score matching when there is concern about group differences in measured covariates and instrumental variable analysis when there is concern about differences in unmeasured covariates.