Non-experimental study designs: The basics and recent advances

Elizabeth A. Stuart, PhD
Bloomberg Professor of American Health
www.elizabethstuart.org
@lizstuartdc

Some causal questions

- How effective currently are the COVID-19 vaccines at reducing severe disease and death?
- What are the effects of moving to virtual instruction on the mental health of middle school students?
- Does air pollution lead to higher mortality?

Why are these questions hard to answer?

- Causal questions inherently involve unobserved quantities:
  - What would have happened under some other state of the world?
- [Note that today I am focusing on estimating causal effects, rather than identifying the causes of effects, which is even harder!]

What is a causal effect?

- Comparison of potential outcomes for THE SAME well defined population:
  - Y(1): Outcome if treated (exposed)
  - Y(0): Outcome if control (not exposed)
- An association compares some outcome in two different groups

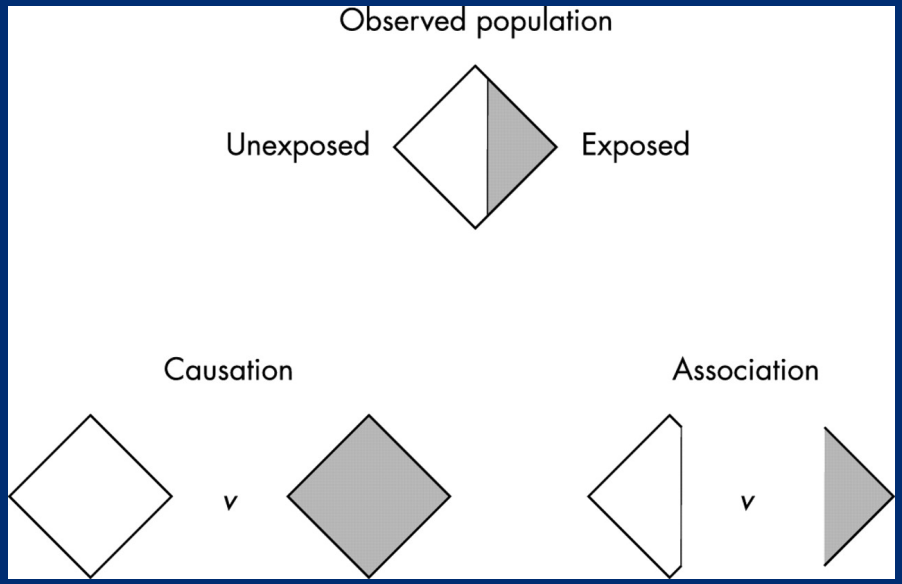


Figure from Hernán (2004): https://jech.bmj.com/content/58/4/265.info

The importance of the estimand vs. the estimator

estimand

estimator

estimate

- It is confusing because we often use data from two different groups to estimate a causal effect
- But important to distinguish the **estimand** – the thing we want to learn about – from the **estimator** -- how we learn about it
- More on this “how do we learn about causal effects” as we go along

Why we can sometimes be led astray

Treatment = Taking aspirin

Outcome = Headache pain 2 hours later

What our data looks like:

Treatment group	Age	Outcome
T	45	12
C	38	10
C	35	8
T	48	14
C	23	12
T	40	9
C	32	17

What our data really is:

Treatment group	Age	Outcome under treatment (Y(1))	Outcome under control (Y(0))	Causal effect
T	45	12	?	?
C	38	?	10	?
C	35	?	8	?
T	48	14	?	?
C	23	?	12	?
T	40	9	?	?
C	32	?	17	?

This is what distinguishes causal inference from standard statistical inference: Trying to learn about things we don't directly observe

So strategies like cross-validation, model checking, model validation mean very different things and can't be directly used

For causal inference we need to rely on smart designs to help us learn about the missing potential outcomes, and thus the causal effect

The value of randomization

- Randomized experiments particularly useful for causal inference
- Treatment and control groups only randomly different from one another, and so the outcomes observed in each group are an excellent proxy (in fact unbiased) for the missing potential outcomes in the other group
    - Note that this was not the case in the little example earlier
    - Treated group somewhat older – may not be able to immediately use their outcomes as an estimate of what would have happened to the (younger) control group had they instead taken aspirin
- Conducting a randomized experiment also forces clarity around the treatment and comparison conditions, outcomes, and timing of measurement – will come back to that!
- (Many of the issues are similar to the considerations between probability and non-probability samples in survey research; Mercer et al., POQ, 2017)

So when we can't randomize…the role of design for non-experimental studies

- Should use the same spirit of design when analyzing non-experimental data, where we just see that some people got the treatment and others the control
- Helps articulate 1) the causal question, and 2) the timing of covariates, exposure, and outcomes

"The planner of an observational study should always ask him[them]self: How would the study be conducted if it were possible to do it by controlled experimentation?" - W. Cochran (1965)

A spectrum of study designs

- We are all familiar with the standard hierarchy of study designs
- Randomized trials seen as best, everything else less good

- What are the study designs that might be useful for estimating causal effects?
- How can we have strong, useful designs that do not sacrifice rigor?

- Key point: In a randomized trial we basically don't need any assumptions to get an unbiased effect estimate

- In (almost) any non-experimental study there will be untestable assumptions we have to make to interpret results as causal

- Key is to assess those assumptions in any particular study

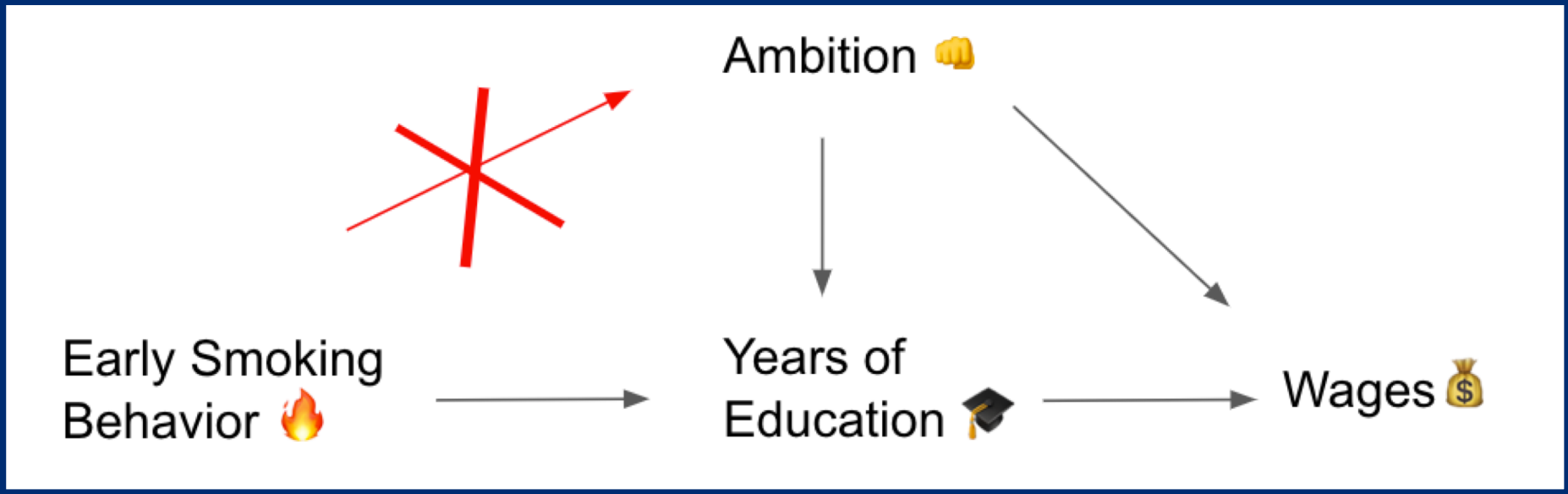
Design 1: Instrumental variables (IV)

- Sometimes may not be randomize access to the treatment itself (e.g., flu shots)
- But could randomize encouragement to take the treatment (e.g., special note from the doctor encouraging flu shot)
- This is known as a "randomized encouragement design" and can be used to estimate the effect of the encouragement OR of the treatment itself
- Known as "instrumental variables," where the encourage is an "instrument" for the treatment of actual interest
- (Involves other assumptions, such as that there is no "direct effect" of the instrument on outcomes)

Naturally occurring IV's

- In some cases you don't actually randomize the instrument but just try to identify a naturally occurring instrument
- Very common in economics
- Fundamentally, need to identify some "instrument" that is related to the treatment of actual interest but does not directly influence outcomes
- Example 1: use random charter school lotteries to study the effects of charter schools
- Example 2: use physician prescribing preference to look at effect of one particular antidepressant vs. another
- (Method can also be used to account for non-compliance in randomized trials)

IV example



Source: Amanda West, Towards Data Science

Consideration of threats to validity

- Key assumptions:
  - Instrument randomized (at least hypothetically)
  - No ”direct effect” of instrument on outcomes
- Randomized encouragement designs can be particularly strong

Design 2: Interrupted time series methods

- These designs useful when we have measures repeated at multiple time points (e.g., monthly) or when only aggregate data is available
- Fundamental idea: Compare what happened after some change happened (the intervention) with what we would have expected to have happened had the intervention not been put into place
- Can be implemented with data from just one unit! e.g., one community that passed a new law
    - Very basic version is a single group pre/post design
    - Best with multiple time periods, to model pre-intervention time trends well
- Design stronger if there is a comparison group with no change, to provide information on general temporal trends before and after
- Analysis approach accounts for correlation of measures across time within a unit

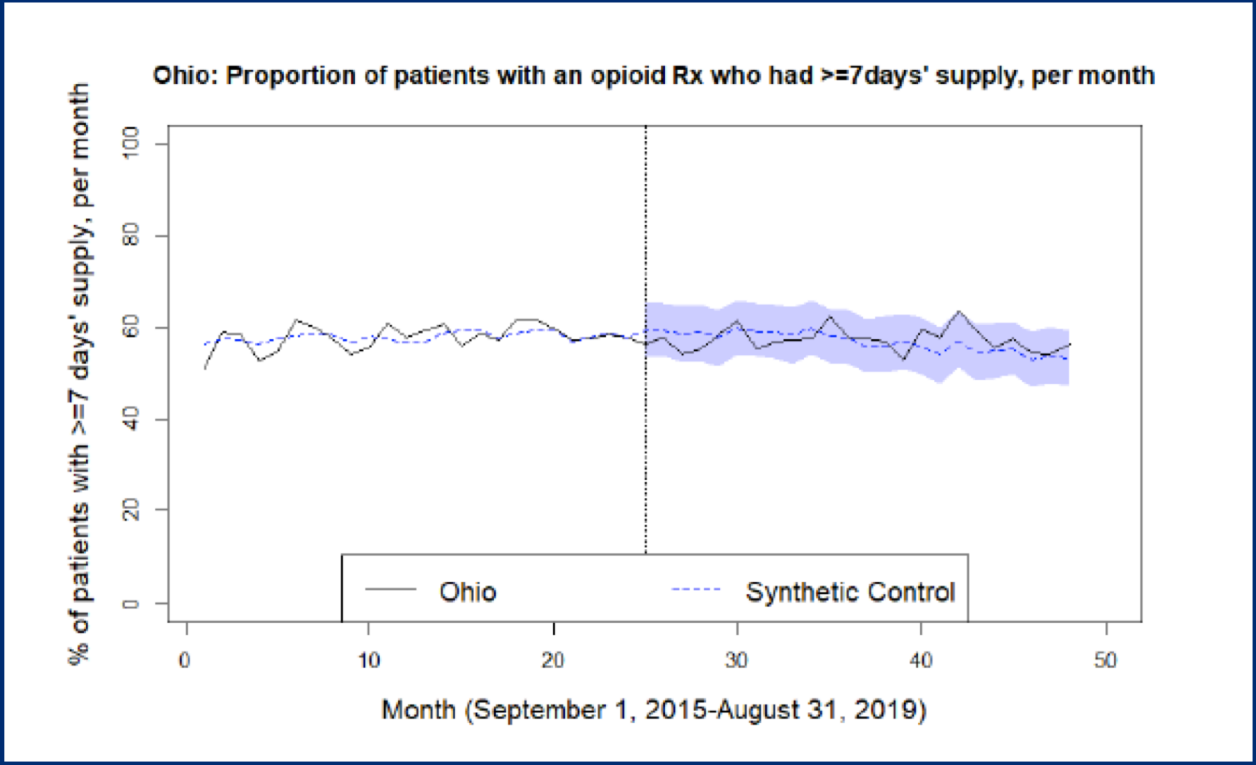
Threats to validity

- Key assumption:
    - Possible to predict the counterfactual trends using baseline data and (if available) the comparison group
    - “Parallel counterfactual trends”
- Comparative interrupted time series generally better, especially if the comparison units had similar trends as the intervention unit during the pre-intervention period
- ITS often fairly easy to implement with existing administrative data
- Should consider model fit and use flexible models relating outcomes to time and covariates

Rapidly evolving methods area

- New understanding of limitations of analysis-based approaches, especially “two way fixed effects” that basically just fits a regression model with location (e.g., state) and time fixed effects, especially in cases of ”staggered implementation”
- Better approaches take more of a design based approach, estimating effects for cohorts one at a time and so having clearly defined pre and post periods
    - Known as “stacked CITS” or “event study” designs
- New synthetic control and augmented synthetic control methods help ensure similar trends in the pre-period

Example: State opioid prescribing laws

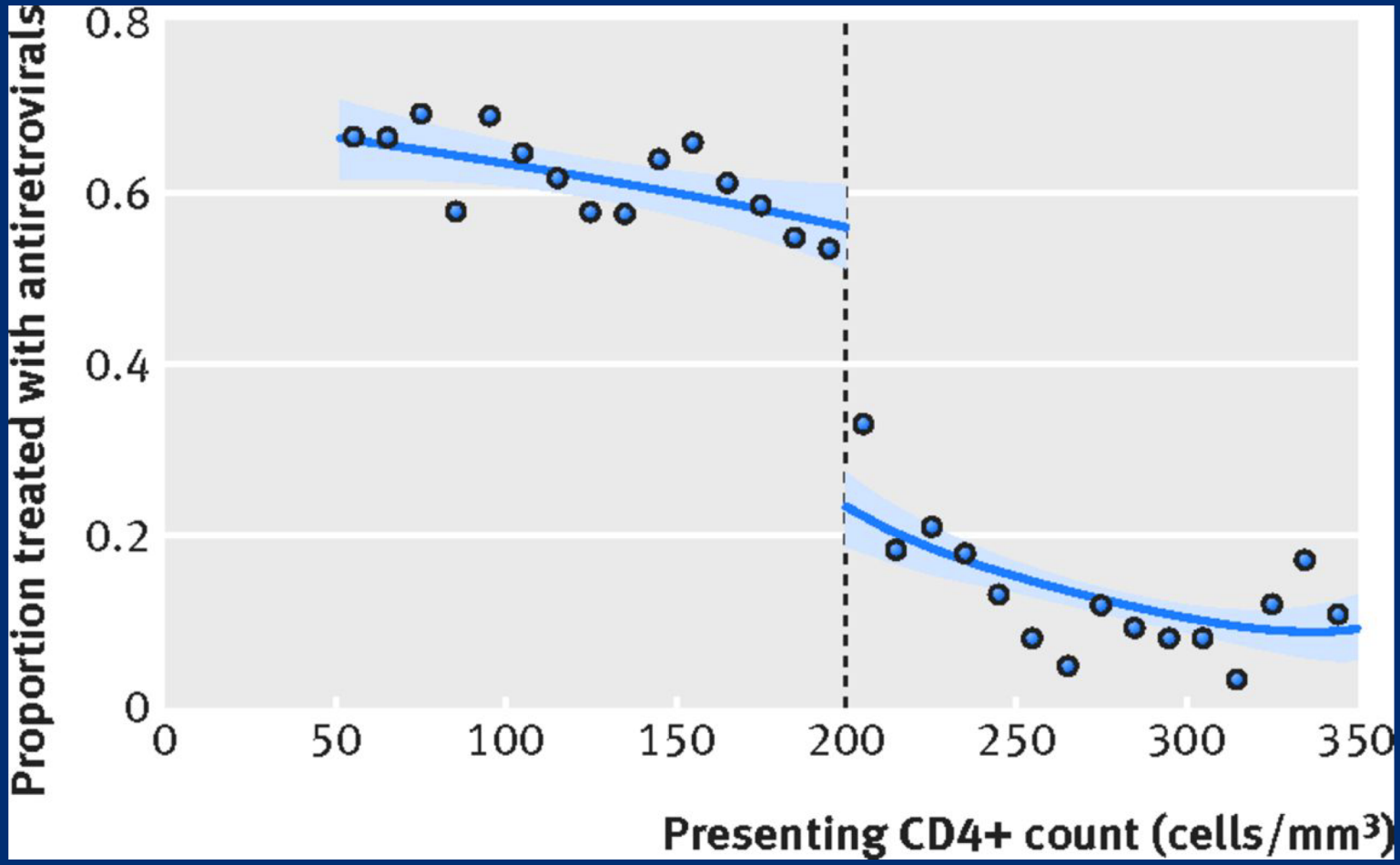


ource: McGinty et al. (Annals of Internal Medicine; 2022)

Design 3: Regression discontinuity

- Useful when program administered on the basis of a cut-off
    - e.g., individuals with a risk score > 10 receive the program; others don't
- Main idea: Like a mini-randomized trial right around the cut-off
- Test whether there is a discontinuity in the outcome at the cut-off; if so, presumption is that it's due to the program

Example



Source: Venkatarmani and Jena (2016)

Threats to validity

- Program has to have been implemented using the cut-off
    - Can have some “fuzziness” but can’t have manipulation of the cut-off (“forcing”) variable
- Best if strong relationship between the cut-off variable and the outcome
- Should check model fit and allow for flexible relationship between cut-off variable and the outcome

Design 4: Comparison group propensity score methods

Estimating the effects of suicide prevention centers in Denmark

Erlangsen et al. (*Lancet Psychiatry,* 2014)

Studying suicide prevention centers

- Suicide prevention programs generally hard to study in a randomized design
- Require large samples, long follow up, and often ethical concerns

- Denmark began rolling out suicide prevention centers around the country in 1992 (now nationwide)

- Causal question: “What is the effect of these centers on the people who go to them, in terms of repeat suicide attempts and death up to 20 years later?”

The data: Danish registries

- Amazing large-scale and comprehensive data on residents of Denmark
- Linked registers: Danish civil register, national registry of patients, psychiatric central registry, registry of causes of death
- Allows longitudinal data on individuals (and their families!!), including extensive social and health information
- Data on individuals 10+ from 1992-2011

A non-design based approach

- Typical analysis approach (especially years ago) would be to just fit a big regression model
- Use data from everyone in Denmark
- Model like: f(Y) ~ Treatment + Covariates
- Interpret coefficient on Treatment as the estimated effect
- Why isn't this great?
  - Not a careful comparison
  - Doesn't anchor time
  - Relies on extrapolation if treatment and control groups dissimilar

Instead will use a design based approach

Design: Propensity score methods

- Perhaps one of the most common non-randomized study designs
- Fundamentally: compare individuals or groups who got the treatment of interest with those who didn't, but do so in a smart way
- Propensity scores help find groups that look similar to one another, but some were treated and others got the comparison condition
    - Theory of propensity scores helps with this

- Better than traditional regression adjustment for confounders because it is less model dependent, and diagnostics more straightforward

- Like traditional regression adjustment it relies on an assumption that there are no unobserved differences between the treatment and comparison groups, after matching on the observed covariates (although can assess sensitivity of results to this assumption)

More details

- The propensity score itself is the predicted probability of receiving the treatment, given the observed covariates
- Often estimated using logistic regression or non-parametric methods such as random forests
- Propensity scores then used to match, weight, or subclassify the individuals

- The goal: Make the treatment and comparison groups look similar with respect to the observed characteristics
- Can check this! Balance measures, which compare covariate distributions between the groups, are crucial

- Other sample equating methods can also be used; propensity scores are just a helpful tool

The treatment and comparison groups in Denmark example

- Treatment group: ~ 6,000 people who went to one of the suicide prevention centers after a suicide attempt

- Comparison group data, used to estimate what would have happened to the treatment group members had they not gone to one of the centers:
    - ~ 60,000 people who had an (index) suicide attempt but then did not go to one of the suicide prevention centers
    - (Wouldn't want to use people without an index suicide attempt; this is part of careful design)

The design

- Use 3:1 propensity score matching to find 3 comparison group individuals for each treated subject
    - Propensity scores estimated using 31 covariates, including demographics, previous suicide attempts, method of attempt, family history, and psychiatric disorders
    - Also require “exact match” on two particularly important confounders: any psychiatric disorder and previous attempts
- Importantly, can check how well this worked!
- Also importantly, done without using the outcome data!

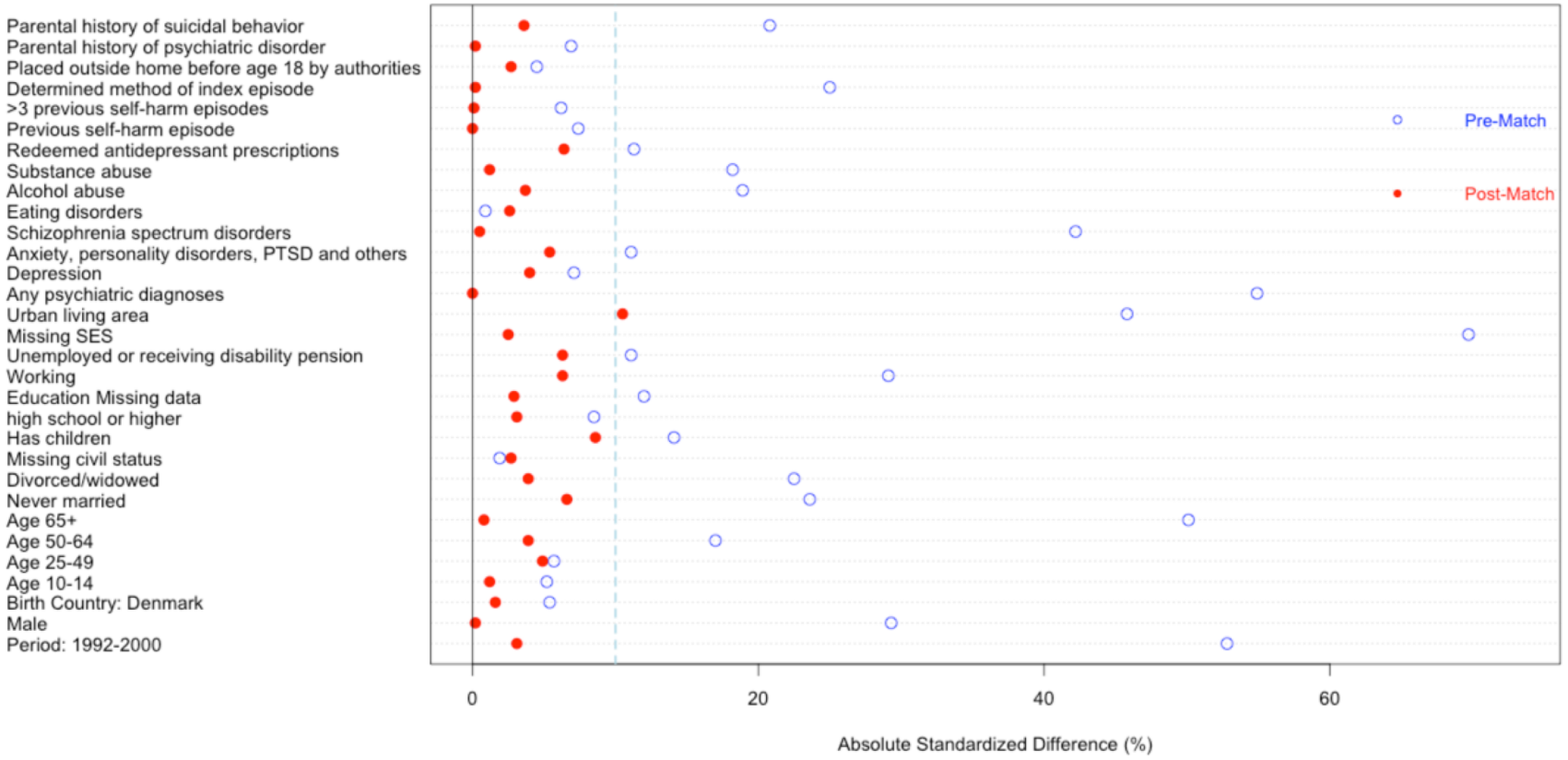
What does the data look like before matching?

Characteristic	Therapy group	Comparison group	Standardized mean difference
Male	31%	45%	0.29
Born in Denmark	90%	91%	0.05
Age 65+	2%	9%	0.50
Has children	39%	46%	0.14
Working	40%	25%	0.29
Any psychiatric diagnosis	72%	48%	0.55
> 3 previous suicide attempts	1.5%	2.3%	0.06

And after?

Characteristic	Therapy group	Matched comparison group	Comparison group	(After) standardized mean difference
Male	31%	31%	45%	0.00
Born in Denmark	90%	90%	91%	0.02
Age 65+	2%	2%	9%	0.01
Has children	39%	43%	46%	0.09
Working	40%	37%	25%	0.06
Any psychiatric diagnosis	72%	72%	48%	0.00
> 3 previous suicide attempts	1.5%	1.5%	2.3%	0.00

Matched groups similar on all 31 covariates



Now can compare outcomes

Outcome	Odds ratio	Confidence interval
Repeat attempt in 5 years	0.80	(0.73, 0.87)
Death by suicide in 5 years	0.74	(0.57, 0.97)
Death from any cause in 5 years	0.66	(0.56, 0.77)

The Achilles Heel: Unobserved confounding

- Still concern there may be an unobserved confounder related to going to one of the Centers and outcomes
- Sensitivity analysis can assess how strong such an unobserved variable would have to be to change study conclusions
- Turn a broad qualitative worry into a more quantitative concrete statement

- For one of the weaker effects (repeated self-harm after 20 years) a binary unobserved confounder with prevalence 0.5 would have to have a 1.8-fold association with participation in the program and a two-fold association with the outcome in order to explain the results

- Substantive experts felt this is unlikely

Threats to validity

- Key assumption: No unobserved confounders, once we adjust for the observed covariates
- So will work best if most (or all) of the confounders are measured
- Important to have an understanding of what led some people to get the treatment and others not (and measure those factors)
- And can do analyses of sensitivity to this assumption
- Can use existing data; often done in a retrospective analysis
- But still want to retain temporal ordering: match on confounders measured before the treatment, and measure outcomes after the treatment

Other considerations

- What if data is from a complex survey?
    - Easiest is to combine propensity score and survey weights
- What about multilevel settings?
    - Hard to answer quickly! Lots of issues, including whether treatment at individual or group level and the confounding structure
- What about missing data?
    - Same lessons as for other settings!

Lessons

- It is possible to use large-scale data to estimate causal effects in non-experimental studies
- Helps to have extensive covariates measured
- Use design elements, such as strategic selection of comparison subjects and approaches to help equate the treatment and comparison subjects on observed characteristics
- And important to acknowledge potential for unobserved confounding

Conclusions

A number of non-randomized designs exist

- A number of non-experimental study designs can be used to estimate causal effects
- Non-randomized designs may be particularly useful for studying effectiveness, given potential “real world" aspect
- May have weaker internal validity than randomized trials, but often stronger external validity
- In any given study need to think through what data is available, and which design fits best (i.e., which assumptions most likely to be satisfied)

Lessons

- Remember that causal inference inherently involves trying to learn about things we don't directly observe
- Think carefully about time and ensure temporal ordering
- Find a devil's advocate/"hostile critic"
- Non-experimental studies will always involve some untestable assumptions
    - [And if someone claims they can test them there must be some other assumption underlying the test!]

Building a body of evidence

- "In conclusion, observational studies are an interesting and challenging field which demands a good deal of humility, since we can claim only to be groping toward the truth." (Cochran, 1972): **No one study will be definitive!**

- **"Cochran's Causal Crossword"** (Rosenbaum, 2015): "To take Cochran's advice seriously is to be skeptical of investigations that derive stout conclusions from slender evidence. It is to be skeptical of grand studies and grand conclusions, the suggestion that a single proposed entry settles a major issue, that consistent completion of the puzzle is inevitable given this one entry, and hence consistent completion is not needed and not worth the effort."

"If only the proponents of big data for causal purposes would take the time to read Cochran's 1972 paper with care!" (Feinberg, 2015)

To learn more…

Fully online short course propensity scores in JHSPH summer institute (also mediation, missing data): http://www.jhsph.edu/departments/mental-health/summer-institute/courses.html

Cochran, W.G. (1972). Observational Studies. Reprinted with commentaries, in *Observational Studies.* https://muse.jhu.edu/article/793409/pdf

Haber, Clarke-Deelder, Salomon, Feller, and Stuart (2020): Policy evaluation in COVID-10: A guide to common design issues. *American Journal of Epidemiology*. https://pubmed.ncbi.nlm.nih.gov/34180960/

Jackson, J.J., Schmid, I., and Stuart, E.A. (2017). Propensity scores in pharmacoepidemiology: Beyond the horizon. Current Epidemiology Reports. Topical collection on pharmacoepidemiology. Published online 6 November 2017. http://link.springer.com/article/10.1007/s40471-017-0131-y.

Stuart, E.A. (2010). Matching Methods for Causal Inference: A review and a look forward. Statistical Science 25(1): 1-21. PMCID: PMC2943670. http://www.ncbi.nlm.nih.gov/pubmed/20871802.

Erlangsen, A., . . . , Stuart, E.A., et al. (2015). Short and long term effects of psychosocial therapy provided to persons after deliberate self-harm: a register-based, nationwide multicentre study using propensity score matching. Lancet Psychiatry.

French, B., and Stuart, E.A. (2020). Study designs and statistical methods for studies of child and adolescent health policies. JAMA Pediatrics 174(10): 925-927