Recent advances in non-experimental comparison group designs

Elizabeth Stuart

Johns Hopkins Bloomberg School of Public Health
Department of Mental Health
Department of Biostatistics
Department of Health Policy and Management
www.biostat.jhsph.edu/∼estuart
estuart@jhsph.edu

Funding thanks to IES R305D150001

September 23, 2016

Outline

Outline

Introduction and Motivation

- Randomized experiments often seen as the gold standard for estimating causal effects (for good reason)
- But some important causal questions can only be answered using non-experimental studies
  - e.g., interventions or risk factors it would be unethical to randomize (child maltreatment, drug use)
  - e.g., not feasible to randomize because intervention widely available (books in the home, online reading program)
  - e.g., can't wait that long to collect outcome data (long term effects of Head Start)
  - e.g., worried that people who participate may not represent the target population (medical trials conducted only in academic medical centers, school-based studies conducted primarily in large districts)

The problem

- Individuals who select one treatment, or who are exposed to some risk factor of interest, likely different from those who don't
  - "Confounding"
  - Hard to separate out differences in outcomes due to these other confounders, vs. due to the treatment of interest

Some non-experimental design options

- Instrumental variables
  - Requires finding an "instrument" that affects the "treatment" received but does not directly affect the outcome
  - Randomized encouragement designs are one (sometimes feasible) type
  - Otherwise have to hope for some naturally occurring instrument (e.g., charter school lotteries)
- Regression discontinuity
  - Requires that treatment administered in a way that used a discontinuity; e.g., students with scores below a threshold got the intervention
- Interrupted time series
  - Useful for interventions implemented at a particular point in time for a particular group (e.g., policy changes), with longitudinal measures before and after
  - *Comparative* interrupted time series better than simple ITS, and then many of the same issues we will talk about here still come up

Comparison group designs as a feasible option

- Comparison groups often one of the most feasible designs
- Main idea: have data on people who got some treatment of interest, find a comparison group of individuals who are similar but did not receive the treatment
- Main strategy: try to ensure that the treatment and comparison groups are as similar as possible on a large set of baseline characteristics
  - Traditionally may have just used regression adjustment to "control for" any differences
  - However, this can lead to model dependence and concerns about model misspecification if the groups are quite different
  - So a large literature has built up that aims to use design to equate the groups before subsequent regression adjustment
  - Propensity scores are one key tool in this design as they help create groups that look similar on a potentially large set of baseline characteristics
  - Big picture common strategies: matching, weighting, subclassification

Talk today

- Will not provide a general history and introduction to these methods
  - Fundamentally, think about them as ways to make treatment and comparison groups "look like" they could have come from a randomized trial
  - (Large literature on the benefits of "emulating" a randomized trial)
- Focus on 3 particular recent advances in this field:
  1. Evidence on when comparison group designs "work"
  2. Moving these designs to more "science" than "art"
  3. Importance of sensitivity analysis to unobserved confounders
- Will also focus on point in time treatments today; more complex settings (e.g., longitudinal treatments) require generalization of these ideas; fundamentals stay the same though

Outline

When the key assumptions hold!

- Key assumption: "unconfounded treatment assignment"
  - No unmeasured confounders: no unobserved differences between treatment and comparison groups, once we have balanced the groups on the observed characteristics
  - Also called "no hidden bias" or "ignorable treatment assignment"
- Also requires assumption that everyone in the study had a non-zero chance of getting treatment or control ("common support")
  - e.g., explicitly exclude people not eligible for the treatment
- So how do we make these assumptions more plausible?

It's all about the covariates!

- Increasing evidence that what matters is what covariates are included, not exactly how the matching/equating is done
- Including a large set of covariates, in particular those related to treatment assignment and outcomes, makes unconfoundedness more likely to be satisfied
- Careful design crucial: what are the important confounders, and do we (or can we) measure them?
- Steiner, Cook, Shadish, and Clark (2010; Psych Methods); Cook, Shadish, and Wong (2008; JPAM); Steiner, Cook, Li, and Clark (2015; JREE)

Figure 2 from Steiner et al. (2010)

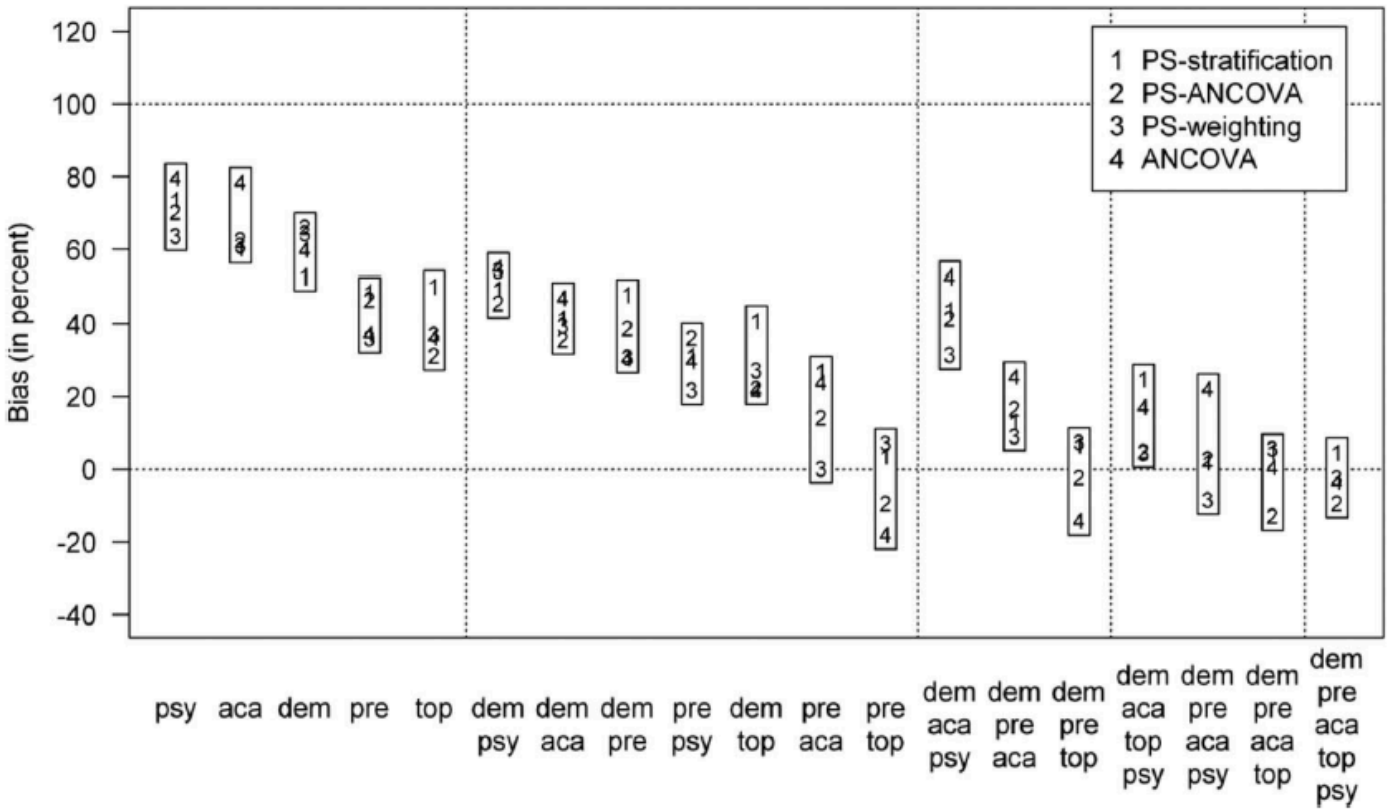

Figure 2. Remaining bias in vocabulary by construct set and analytic method (in the order of average bias). Construct sets are composed of demographics (dem), proxy-pretests (pre), prior academic achievement (aca), topic preference (top), and psychological predisposition (psy). PS = propensity score; ANCOVA = analysis of covariance.

Lessons from this literature

- Think carefully about comparison group selection ("clever design")
- Use a large set of variables (not just demographics; also include, e.g., pretest measures of the outcome)
- Select comparison group carefully (e.g., from same geographic area)
- Measure variables in same ways across treatment and comparison groups
- (Using large national datasets usually not as effective)
- Have good understanding of the treatment selection process (importance of the assignment mechanism!)
- Have large sample sizes in the comparison groups: easier to get good balance
- (Also should not adjust for/match on post-treatment variables)

Outline

More automated methods

- Traditionally the use of methods such as propensity scores has involved a fair amount of "art"
- Goal: create groups that look similar on the observed covariates (covariate "balance")
- To get there, try different estimation techniques, equating methods, interaction terms, etc., and pick the one that gives the best covariate balance
- New methods aim to remove some of this iteration, in two ways:
  - Get balance on the covariates themselves directly (not necessarily using the propensity score)
  - Estimate and use the propensity score in an automated way to get good balance

Getting direct balance

- Some methods aim to directly balance the covariates, not "through" the propensity score
- Coarsened exact matching (CEM; King et al.; cem R package): http://gking.harvard.edu/cem
  - Essentially, exact matching on coarsened (categorized) covariates
  - Trade off between number of matches and closeness of matching
  - Works well for easily categorized variables (like high school degree or not); less clear for truly continuous ones (like age)
- Mixed integer programming and "fine balance" (Zubizarreta, 2012; mismatch R package)
  - Sort of like exact matching on a few variables
  - But instead of getting individual-level exact matches, matches the distributions exactly across the matched samples

Zubizarreta et al. (2015), Table 2

TABLE 2. Distributions of Sex, Age, Self-Rated Health, and Housing Quality Before the Earthquake

	After matching	
	No. exposed (n = 2,520)	**No. controls (n = 2,520)**
Sex		
Men	831	831
Women	1,689	1,689
Age (years)		
15–24	195	195
25–34	412	412
35–44	561	561
45–54	474	474
55–64	406	406
65+	472	472
Ethnic group		
Indigenous	210	210
Nonindigenous	2,310	2,310

More automated propensity score methods

- Other methods aim to automate the propensity score estimation itself, to estimate the propensity score in a way that optimizes balance
- Most popular: Covariate Balancing Propensity Score (CBPS; Imai and Ratkovic): http://imai.princeton.edu/research/CBPS.html
  - Doesn't simply maximize the likelihood; also has a balance constraint that it jointly maximizes
  - Benefit of this is that it maintains the nice theoretical properties of the propensity score, but also more directly targets balance
- Genetic matching (Sekhon et al.) another version of this
- One drawback: many of these optimize a particular balance measure and may not optimize others
  - Need more work to determine the best balance measures to optimize

Outline

What if we don't believe unconfoundedness?

- Sensitivity analyses can be done to assess how sensitive results are to an unobserved confounder
- Ask the question: How strongly related to treatment assignment and outcome would such a factor have to be in order to change study conclusions?
- Based originally on analysis by Cornfield showing that association between smoking and lung cancer most likely actually causal
- Methods now extended by Rosenbaum, VanderWeele, others

Example: Effects of psychosocial therapy on repeat suicide attempts

- Erlangsen et al. (2014) used Danish registry data to estimate effect of suicide prevention centers on
- Concern that there may be an unobserved variable related to participation and outcomes
- Sensitivity analysis can assess how strong such an unobserved variable would have to be to change study conclusions
  - Used approach by VanderWeele and Arah (see Liu et al., 2013)
- 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

Outline

Strengths and limitations of non-experimental comparison group designs

- Strengths
  - Often feasible
  - Relatively easy to describe and understand
  - Idea of replicating a randomized trial: no use of outcome data in setting up the design
- Limitations
  - Relies on having high-quality data
  - Common measures across treatment and control groups, and important confounders measured
  - Helps to have a good understanding of the treatment selection process, which is rare (opportunity for combining qualitative and quantitative work??)

Conclusions

- Many research questions require non-experimental designs
- When using non-experimental comparison group designs clever design helps
- General lessons:
  - Measure as many confounders as possible; try to have an understanding of the treatment selection process
  - Try to get as good covariate balance on the observed covariates as possible
  - Assess sensitivity to key assumption of no unmeasured confounders
- Also lots of complications and extensions: multiple treatment levels, time-varying treatments, missing outcomes, . . .

"It's been fifty years now. I guess you can't compare apples to oranges."

And remember . . .

"With better data, fewer assumptions are needed."
- Rubin (2005, p. 324)

"You can't fix by analysis what you bungled by design."
- Light, Singer and Willett (1990, p. v)

References

- Online class at JHSPH: 140.664 ("Causal inference in medicine and public health")
- One-day short course on propensity scores in JHSPH summer institute (in-person and online!): http://www.jhsph.edu/departments/mental-health/summer-institute/courses.html
- Erlangsen, A., . . . , Stuart, E.A., et al. (2014). 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*.
- Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15(3): 199-236. http://gking.harvard.edu/matchp.pdf.
- Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation.*Health Services & Outcomes Research Methodology* 2, 169-188.
- Stuart, E.A. (2010). Matching Methods for Causal Inference: A review and a look forward. *Statistical Science* 25(1): 1-21
- Liu, W., Kuramoto, S.K., and Stuart, E.A. (2013). An Introduction to Sensitivity Analysis for Unobserved Confounding in Non-Experimental Prevention Research. *Prevention Science* 14(6): 570-580. PMCID: 3800481.