Getting ready for propensity score methods: Designing non-experimental studies and selecting comparison groups

Elizabeth A. Stuart
Johns Hopkins Bloomberg School of Public Health
Departments of Mental Health and Biostatistics

estuart@jhsph.edu
www.biostat.jhsph.edu/∼estuart

Funding thanks to NIMH 1K25MH83646-01

September 6, 2012

Outline

Outline

Where does the data come from?

- When written up, the source of the data often taken as a given
- "We compared treated and comparison individuals . . ."
- But where the data comes from, and what it contains, often the most important step
- Increasing availability of public use, large-scale datasets; can it be useful?
- Will outline some of the things to think about at this initial design stage of a study

The context for my talk

- Have data on a treated/exposed group
- Want to estimate the effect of that treatment, relative to some comparison condition (which may not yet be fully defined)
  - Causal effect is comparison of potential outcomes between two treatment conditions
  - Can't just look, e.g., at change over time in treatment group
- Don't have a clear comparison group identified, also don't have an obvious instrumental variable
- So hope to use a propensity score method to estimate the treatment effect
- In that case, what do you need to think about up front?
- (Lots of connections with ideas discussed by Dr. Cook and Dr. Thoemmes)

Careful design of observational studies

- Broader discussions of threats to validity, careful design of observational studies
- Think creatively about comparisons, make crisp comparisons
- e.g., Kuramoto et al. (2011): In estimating the effect of having a parent die from suicide, compare to people whose parents died in an accident
- e.g., Rosenbaum (2009): When looking at effects of road features on accidents, compare conditions where accident happened with the conditions for the same car 1 mile before the accident
- Shadish, Cook, and Campbell (2002), Rosenbaum (2009, 2011; "design sensitivity")

Outline

Understanding the treatment assignment process

- To have confidence in causal statements, need to understand the assignment mechanism
- What led some people to get the treatment and others not? Who made the decision, and what was it based on?
- And do we observe those characteristics? Can we adjust for them?
- Propensity score approaches will be more appropriate if we observe the factors that went into this decision
- Do qualitative interviews with "decision makers": Great setting for a mixed methods study?

Comparability of treated and comparison subjects

- Want to identify comparison subjects likely to be similar to treated subjects
- Ideally on observed and unobserved characteristics
- Careful selection can reduce worries about unobserved confounding
- e.g., from same geographic area
- e.g., satisfy eligibility criteria
- (Think back to car accident example)

Availability of appropriate measures

- Most propensity score methods rely on assumption of no unmeasured confounders
- Given the observed characteristics, there are no unobserved variables related to treatment and outcome
- Methods won't work very well if all you have are basic demographics
- Most important: pre-treatment measures of the outcomes
- And of course measures need to be the same across the treatment and comparison groups
  - Having lots of intensive data on just the treatment group won't help; need it on comparison subjects too
- Steiner et al. (2010)

Outline

Multiple control groups and/or biases of a known direction

- Consider a setting where treated individuals must meet some eligibility criteria
- Worry that those not eligible likely quite different from those who are (e.g., less disadvantaged)
- But those who are eligible but don't participate may be quite different from those who do participate (e.g., less motivated)
- So what to do?
- Potential solution: multiple comparison groups
- Works when you can formulate hypotheses about the ordering of outcomes (i.e., hypotheses about the directions of any biases)
- e.g., treatment group should perform better than those who were eligible but didn't participate, but worse (or the same as) those who weren't eligible

Rosenbaum (2009): From Bernanke (1995), Did adherence to the gold standard lengthen the Great Depression?

"...countries that left gold relatively early had largely recovered from the Depression, while Gold Block countries remained at low levels of output and employment ... Of course, in practice the decision about whether to leave the gold standard was endogenous to a degree ... [Any] bias created by endogeneity of the decision to leave gold would appear to go the wrong way ... The presumption is that economically weaker countries ... would be the first to devalue or abandon gold. Yet the evidence is that countries leaving gold recovered substantially more rapidly and vigorously than those who did not. Hence, any correction for endogeniety ... should tend to strengthen the association of economics expansion and the abandonment of gold."

SDDAP: Using internal and external controls

- Estimating the effect of a school-wide dropout prevention program (SDDAP)
- Have kids in 5 treated schools, as well as data on kids from paired comparison schools in the same communities
- The catch: The kids in the treated and their comparison schools don't always look similar on student characteristics
- Trade off between good matches on individual-level characteristics vs. on local area-level characteristics

- Rubin and Stuart (2008) use a combination of local and non-local matches
  - Get a local match when they look good, but get a non-local match if no good local match
- Can also adjust for local/non-local difference by comparing local and non-local comparison students
- Can also estimate optimal number of local matches
  - Solution depends on how close the local groups are and on relative importance of individual level covariates and area indicator in terms of outcome

Genzyme: Using historical data

- FDA submission for approval of Fabrazyme
- Fabry disease very rare
- The goal: Extend double-blind randomized trial to look at longer-term effects
- The challenge: No existing treatment, so for ethical reasons drug became open-label after showing initial promise
- The solution: Use historical patients as comparison subjects (using historical patient registry)

Propensity score matching with historical data

- Utilize two data sources:
  - Group randomized to receive Fabrazyme in the double blind trial
  - Historical control data set with information on historical patients with Fabry disease (primarily clinical data)
- Use propensity score matching to select historical control subjects who look like the treatment group at the time they started Fabrazyme
  - Restrict historical subjects to those that would have met inclusion/exclusion criteria for trial
  - Matching done on variables selected by clinicians as being important for prognosis
- Compare long term outcomes of the treatment group with their matched comparison subjects
- Worked well here in part because treatment for Fabry (and prognosis) had not otherwise changed much over time and measures were the same in the two groups

A related complication: defining baseline

- When longitudinal data available on comparison subjects, may not be clear what to use as their "baseline"
- Need some way of determining which time points correspond to "baseline" and which to "outcomes"
- Would like to identify the point in time they *could have* started treatment but didn't
- Sometimes can use calendar time as an anchor
- In Genzyme example, selected the point in time each comparison subject looked the most like someone in the trial

Utilizing "natural experiments"

- Of course sometimes we may be able to use a "natural experiment"
- Some randomization occurring by chance in nature
- e.g., use Dutch famine of 1944-1945 to examine effects of malnutrition in utero on later mental performance (Rosenbaum, 2009)
- Big literature on this; won't focus on it here

Outline

Rosenbaum, 2005: Some overall lessons in design: Try to have...

- Covariates and outcomes available for treated and control groups, ideally with temporal ordering
- "Haphazard treatment assignment rather than self-selection" (e.g., Maimonides rule, car crashes where not at fault)
- "Special populations offering reduced self-selection" (e.g., violence rates among felons)
- "Biases of known direction" (e.g., a group known to be less disabled than the treated group)
- "An abrupt start to intense treatments" (e.g., car crashes)
- "Additional structural features in quasi-experiments intended to provide information about hidden biases."

Lessons

- Good selection of comparison groups requires creativity
- Don't just take the "easy" data
  - Make sure there are comparable measures of the important confounders
  - Think about finding comparison group data that helps control for unobserved variables
  - Think about whether factors like geography, calendar time matter
- Make a story
  - Compare multiple groups
  - Examine multiple outcomes, some that should/shouldn't be affected by treatment ("zero checks")
  - Elaborate theories
- Value transparency
- Think of a "hostile critic"
- Replication: In multiple settings, slightly different ways of looking at the question

References

- Imai, K., King, G., and Stuart, E.A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 171: 481-502.
- Rosenbaum, P.R. (2009). *Design of Observational Studies.* Springer Verlag, New York, NY.
- Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company.
- Steiner, P.M., Cook, T.D., Shadish, W.R., and Clark, M.H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods* 15(3): 250-267.
- Stuart, E.A. (2010). Matching Methods for Causal Inference: A review and a look forward. Statistical Science 25(1): 1-21.
- Stuart, E.A. and Rubin, D.B. (2008). Matching with multiple control groups and adjusting for group differences. *Journal of Educational and Behavioral Statistics* 33(3): 279-306.
- Fabrazyme example: http://www.fda.gov/ohrms/dockets/ac/03/briefing/3917B1_01_Genzyme.pdf