Analysis methods for improved external validity

Elizabeth Stuart

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

Funding thanks to K25 MH083846, NSF DRL-1335843

April 24, 2014

Outline

1. The scenario
2. Analysis methods
   - Broad assessment of similarity
   - Overview of analysis strategies
   - More on weighting
3. What data do we need for each of these?
4. Bigger picture . . .

Outline

The scenario considered

- Want to make statements about the likely effects of an intervention in a target population
- Have a single study, already conducted
- Also have some data on that target population
  - (And assumes that target population is well defined)
- What can you do?

This talk will . . .

- First talk about existing study analysis strategies
- And then some caveats and concerns
- Note: Focused on assessing and enhancing external validity with respect to the characteristics of trial and population subjects
- Lots of other threats to validity as well: scale-up problems, implementation, different settings

Outline

Outline

Comparison of trial and population

- People often show a "Table 1" that shows differences on a set of characteristics
  - Particularly helpful if can show similarity on outcomes between control group and the population
  - Weisberg et al. (2009), Greenhouse et al. (2008)
- Re-AIM framework (Green and Glasgow, 2006) has specific metrics
  - Attrition, participation rates, quality of implementation
- But often fairly ad-hoc; hard to know what to make of it all

Outline

For settings with multiple studies

- Meta-analysis
  - Quantitatively and formally combine estimates to come up with overall summary of results
  - Often done only with randomized trials
  - Of course if all of the trials sampled the same non-representative population, may not help!
- Cross-design synthesis/research synthesis
  - More general: combine results from multiple studies, including a variety of types
  - Model effect estimates as a function of study characteristics
  - Can include prior distributions on relative biases (Turner et al. 2009)
  - Assess bias due to strict inclusion/exclusion criteria (Pressler & Kaizar, 2013)

For single studies

- Post-stratification
  - Estimate effects separately for subgroups, re-weight those effects to match population distributions
  - e.g., if sample 20/80 male/female but population 50/50, equally weight male- and female-specific effect estimates
  - Only requires joint distributions of the set of key confounders
  - Hard to do with continuous covariates or many categories
- Reweighting
  - Model probability of participation in trial
  - Reweight trial members to weight up to full population using inverse probability of participation weights (like survey sampling or non-response weights)
  - (Generally) requires individual-level data for population
  - Shadish, Cook, and Campbell (2002); Cole and Stuart (2009); Stuart et al. (2009); Tipton (2013); O'Muircheartaigh and Hedges (2014)

- Flexible regression models of the outcome
  - Another strategy is to model the outcome as a function of treatment status, other covariates
  - Then predict outcomes under treatment and control for individuals in the population
  - Primary drawback of basic models like this is model dependence
    - If trial sample and population very different on the covariates, results will rely on extrapolation from sample to population
    - Exactly the same problem that propensity scores try to deal with in non-experimental studies
  - However, flexible regression models can work well
    - e.g., BART: Bayesian Additive Regression Trees
    - Hill (2010) provides evidence of ability of BART to predict population treatment effects

Outline

Weighting

- Like a smoothed version of post-stratification
  - Can work with larger numbers of covariates
  - But relies on propensity score-type model to do that smoothing
- Same idea as non-response weights in surveys, propensity score weights in non-experimental studies
- Case study from Cole & Stuart (2010)

The ACTG Trial

- Examined highly active antiretroviral (HAART) therapy for HIV compared to standard combination therapy
- 577 US HIV+ adults randomized to treatment, 579 to control
- Intent-to-treat analysis: Hazard ratio of 0.51 (95% CI: 0.33, 0.77)
- Evidence of differential effects by age, sex, race

The target population

- What would the effects of HAART be if implemented nationwide?
- US estimates of the number of people infected with HIV in 2006 (CDC, 2008)
- HIV incidence was estimated using a statistical approach with adjustment for testing frequency and extrapolated to the US
- Sample and population differ on age, sex, and race

Comparing trial and population

Characteristic	Trial	2006 US Population
Age groups		
13-29	9%	34%
30-39	45%	31%
40-49	34%	25%
50-75	13%	10%
Male	83%	73%
Race		
White, non-Hisp	54%	36%
Black, non-Hisp	28%	46%
Hispanic	18%	18%
CD4 count, cells/mm^3	75 (33, 137)	NA
N	1156	54,220

Weighting the trial to the population

- Fit model of selection into the trial as a function of characteristics observed in trial and population
- Weight the trial subjects up to the population, using weights that are the inverse of their probability of being in the trial
  - Like survey non-response adjustment

	Hazard ratio	95% CI
Crude trial results	0.51	0.33, 0.77
Age weighted	0.68	0.39, 1.17
Sex weighted	0.53	0.34, 0.82
Race weighted	0.46	0.29, 0.72
Age-sex-race weighted	0.57	0.33, 1.00

Placebo checks

- Can also use the weighting as a diagnostic
- In two ways:
  - Selection probability itself captures differences between the groups in a scalar summary (like a propensity score)
  - Weighted control group mean should match the population outcome mean if the control conditions are the same ("placebo check")
- In HAART case, if we had mortality information in the population, could see if weighted mortality rate among control group matched the population mortality rate (assuming no treatment in the population)
- If placebo check fails, may indicate unobserved differences between the groups
- Hartman et al., 2013; Stuart et al., 2011

Outline

Data needed for each method

	Trial Data	Popn. Covariates	Popn. Outcomes	Popn. Treatment
Weighting	X	X*	Helps	
Regression, BART	X	X	Helps	
Comb. Exp and Non-Exp	X	X	X	X

* Individual level data not necessarily needed

Outline

What do we need to assess and enhance external validity?

- Information on the factors that influence treatment effect heterogeneity
- Information on the factors that influence participation
- Data on all of these factors in the trial and the population
  - Not very helpful if these factors not observed in the population
- Methods that allow for the differences between trial and population on these factors

Data a primary limiting factor

- Right now we have very little information on factors that influence effects or participation in trials
- Often hard to find population data
- Even harder to find population data that has the same measures as trial of interest
  - Case study of Head Start Impact Study and REDI sample
- Better to have more things observed
  - Again, analogies with non-experimental studies

Recommendations

- Get better information on treatment effect heterogeneity
  - Better analyses of existing trials
  - Meta-analysis of existing trials
  - Theoretical models for the interventions
- Get better information on factors that influence participation in trials
  - We know almost nothing at this point
- Standardize measures
  - At least make it more feasible to combine trial and population information
- More research on the methods, and understanding when they work (and when they don't)
  - Many rely on strong assumptions, and preliminary evidence that they can be quite sensitive to those assumptions

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

- Cole, S.R. and Stuart, E.A. (2010). Generalizing evidence from randomized clinical trials to target populations: the ACTG-320 trial. *American Journal of Epidemiology* 172: 107-115.
- 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.
- Olsen, R., Bell, S., Orr, L., and Stuart, E.A. (2013). External Validity in Policy Evaluations that Choose Sites Purposively. *Journal of Policy Analysis and Management* 32(1): 107-121.
- Stuart, E.A., Cole, S.R., Bradshaw, C.P., and Leaf, P.J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *The Journal of the Royal Statistical Society, Series A* 174(2): 369-386.
- Stuart, E.A. (in press). Generalizability of clinical trial results. Forthcoming in *Methods in Comparative Effectiveness Research.* Edited by Constantine Gatsonis and Sally C. Morton.