Dealing with observed and unobserved effect moderators when estimating population average treatment effects

Elizabeth Stuart
Associate Dean for Education
Bloomberg Professor of American Health
www.elizabethstuart.org, @lizstuartdc

October 9, 2020

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Outline

1. Introduction, context, and framework
2. How much do we need to worry about this?
3. How could we estimate the PATE if we wanted to?
4. But what about unobserved moderators?
5. Conclusions

Outline

Making research results relevant: A range of policy or practice questions

- A health insurer may be deciding whether or not to approve payment for a new treatment for back pain
- Interest in predicting overall population impacts of a broad public health media campaign around not switching car seats to forward facing until a child is 12 months old
- A physician practice may be deciding whether training providers in a new intervention would be cost effective across their population
- A health care system might want to know whether giving convalescent plasma to all individuals recently diagnosed with COVID-19 would be useful

From individual to population effects

- All of these reflect a “population" average treatment effect
  - e.g., across individuals in a population, does this intervention work “on average"?
  - This population could be fairly narrow, or quite broad
- There may actually be underlying treatment effect heterogeneity
  - e.g., stronger effects for some individuals
  - Lots of interest in tailoring treatments for individuals; not my focus today
- For policy questions that motivate today’s talk, desire an overall average effect for a well-defined target population

At this point, relatively little attention to how well results from a given study might carry over to a relevant target population

Some recent methods have been developed to estimate population effects from a randomized trial sample

This talk will discuss those, as well as the common data complications that arise, especially that often we have limited data observed!

Generalization more broadly

- There are lots of reasons why results from randomized trials may not generalize
  - Scale-up problems, different contexts, different implementation
- Today will focus on differences due to differences between a trial sample and a population in characteristics that moderate treatment effects
- Note 1: Some people use term “transportability;" today I use generalizability but same ideas hold
- Note 2: Similar issues arising in debates about non-probability samples in survey world (Mercer et al., 2017)

Formalizing trade-offs of different designs (Imai, King, & Stuart, 2008)

- Interested in the effect of some treatment, T, on an outcome, Y, in some target population
  - Population average treatment effect (PATE)

$$\text{PATE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$$

- Simple estimate of PATE is just a difference in means of the outcome between the observed treated and control groups

$$D = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n_0} Y_i(0)$$

- Bias in estimated treatment effect: Δ=PATE - D
- How do different design elements affect the size of Δ?

Decomposition of Δ

Different study designs entail trade-offs between sources of bias

$$\Delta = (\Delta_{SX} + \Delta_{SU}) + (\Delta_{TX} + \Delta_{TU})$$

- Decompose Δ into 4 parts, due to:
  - Sample selection bias (S), Treatment selection bias (T)
  - Observed variables (X), Unobserved variables (U)

Examples of trade-offs of different designs

Study design	Sample Selection Bias		Treatment Selection Bias	
	Δ_{SX}	Δ_{SU}	Δ_{TX}	Δ_{TU}
Ideal experiment	0	0	0	0
Typical experiment	Big	Big	0	0
Typical non-exp study	Small	Small	Big	Big
Well-done non-exp study	Small	Small	Small	?

Thinking through these terms

- Much of my work focuses on methods to reduce Δ_{TX} in non-experimental studies
  - And some on how to assess sensitivity of results to Δ_{TU}
  - Propensity scores, matching methods more generally, sensitivity to unobserved confounders
- Today focusing on Δ_{SX} and Δ_{SU}
  - How big are they?
  - Can we use data and statistical methods to make them smaller, or do sensitivity analyses around Δ_{SU}?

Outline

When will there be external validity bias?

- Intuitively (and formally), there will be bias if participation in trial associated with impacts
- In particular, external validity bias a function of:
  - Variation in probabilities of participation in trial
  - Variation in treatment effects
  - Correlation between those two things
- If any of these factors is 0, then no bias. But otherwise it exists.
- Unfortunately very hard to estimate these quantities in populations (impossible?)
- Formalized in Olsen et al. (2013) and Cole and Stuart (2010)

What about in real data?

- Unfortunately almost no empirical evidence yet on actual size
  - Need good estimate of treatment effect in population of interest, and estimate in samples that would have participated in a trial
- Wisniewski et al. (2009) subset large pragmatic trial (STAR*D) to those eligible for a more narrow (and more typical) efficacy trial: found better outcomes and larger impacts in efficacy sample
- Bell et al. (2017) compared impact estimates in samples actually selected for 11 US government-funded educational evaluations to “true" population impact (estimated using regression discontinuity design); found bias of about 0.1 standard deviations
- External validity bias may be of similar magnitude as internal validity bias in (not well done) non-experimental studies

So instead we are left trying to get evidence on the pieces

- Selection into trials, and whether probabilities of participation vary across individuals
  - But often this not documented very well
  - Especially how people ended up in the trials
- Treatment effect heterogeneity
  - But this hard to assess in trials: almost never powered to detect subgroup effects
  - Growing area of research, and can help inform questions about generalizability
- And understand whether there are factors that influence both (and then try to observe as many of them as possible!)

Differences between sample and population

- At least some attention to how samples are selected or how they differ from target populations
- Susukida et al. (2016, 2017): individuals in the NIDA Clinical Trials Network trials more female, older, and more educated than individuals seeking treatment nationwide
- Humphreys et al. (2005): exclusion criteria in RCTs of alcohol use treatments exclude 20-33% of patients
- Okuda et al. (2010): eligibility criteria in cannabis treatment RCTs would exclude 80% of patients
- Stuart et al. (2017): Districts that participate in large-scale rigorous evaluations of education programs larger, more urban, and lower performing than typical districts nationwide

Outline

Improving external validity through trial design

- Random sampling (great, but rare)
- Purposive sampling (not formally representative)
- Practical clinical trials (potentially useful, but expensive and still lacks formal representativeness)

- Main idea: select subjects for trial in a particular way

Analysis possibilities for settings with multiple studies

- Meta-analysis
  - Quantitatively and formally combine estimates to come up with overall summary of results
  - 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
  - Can include prior distributions on relative biases (Turner et al. 2009)
  - Assess bias due to strict inclusion/exclusion criteria (Pressler & Kaizar, 2013)

Analysis possibilities for single studies

- Post-stratification
  - Estimate effects separately for subgroups, re-weight those effects to match population distributions
- Weighting
  - Model probability of participation in trial, weight trial members to weight up to full population using inverse probability of participation weights (like survey sampling or non-response weights)
  - 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
  - Model the outcome as a (flexible) function of treatment status, other covariates, predict outcomes under treatment and control for individuals in the population
  - e.g., BART: Bayesian Additive Regression Trees (Hill, 2010; Kern et al., 2016), TMLE (Rudolph et al., 2014)

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
- 33/577 and 63/579 endpoints (AIDS/death) during 52-week follow-up
- Intent-to-treat analysis: Hazard ratio of 0.51 (95% CI: 0.33, 0.77)

Cole & Stuart (2010)

The target population

- Don’t necessarily just care about people in trial
- What would the effects of the treatment 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
- We do not have individual-level data but do have the joint distribution (i.e., cross-classification) of important select characteristics; namely, sex, race and age groups
    - Use that to generate pseudo-population representing the target population

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

Inverse probability of selection weighting

- Weight the trial subjects up to the population
- Each subject in trial receives weight $w_i = \frac{1}{P(S_i=1|X)}$
  - (Inverse of their probability of being in the trial)
- Use those weights when calculating means or running regressions
- Related to inverse probability of treatment weighting, Horvitz-Thompson estimation in surveys

Sources of bias in PATE

- Age, race, and gender all significantly associated with membership in trial
  - People in trial more likely to be older, male, white or hispanic (not Black)
- These characteristics also moderate effects in the trial
  - Largest effects for those 30-39, males, and Black individuals

Estimated population effects

	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

- CI's longer for weighted results
- Effects generally somewhat attenuated, except for weighting only by race

Outline

Unobserved effect moderators likely the rule, not the exception

- Key assumption is that all effect moderators related to sample selection observed and adjusted for
  - Kern et al. (2016) showed that methods for estimating PATE all worked quite well when this assumption satisfied, but all bets off when violated
- Rare to have extensive data overlap between sample and population
  - Stuart and Rhodes (2017) found only 7 variables in common between sample and population in early childhood setting
- Need sensitivity analyses to assess robustness of results to violation of this assumption
  - Analogous to analyses of sensitivity to unobserved confounding in non-experimental studies
  - Alternative: bounds (Chan, 2017)

Notation

A: treatment (0,1), randomized in the trial
Y: outcome (observed only in the trial)
Y^a: potential outcome under treatment a, $a = 0, 1$

$S = 1$: trial participation
$P = 1$: target population membership

$$\text{SATE} = \text{E}[Y^1 - Y^0 | S = 1]$$
$$\text{TATE} = \text{E}[Y^1 - Y^0 | P = 1]$$

X: non-effect-modifying covariates
Z: effect modifiers, observed in trial and target population

either V: effect modifier observed in trial, not population

or U: fully unobserved effect modifier

Key Assumptions

A1 *Internal validity of the trial:* conditional ignorability of treatment assignment, positivity, no interference, etc.

A2 *Across-setting treatment variation irrelevance*

A3 *Positivity: everyone in population has non-zero probability of being in trial sample*

A4 *Conditional sample ignorability for treatment effects:* $[Y^1 - Y^0] \perp \{S, P\} \mid Z, V, (S = 1 \text{ or } P = 1)$

A5 *No measurement error:* X, Z are measured the same way in trial and target population, and measured without error

A6 *Additive potential outcomes model:*

$$\mathrm{E}[Y_i^a] = \beta_0 + \beta_x X_i + \beta_z Z_i + \beta_v V_i + \beta_a a + \beta_{za} Z_i a + \beta_{va} V_i a$$

For a partially unobserved effect modifier (V case)

$$\text{TATE} = \beta_a + \beta_{za}\text{E}[Z \mid P = 1] + \beta_{va}\text{E}[V \mid P = 1]$$

Two options:

1. Outcome-model-based sensitivity analysis
   i. obtain estimate for $\text{E}[Z \mid P = 1]$ and specify range for $\text{E}[V \mid P = 1]$
   ii. estimate $\beta_a, \beta_{za}, \beta_{va}$ using trial data
   iii. combine

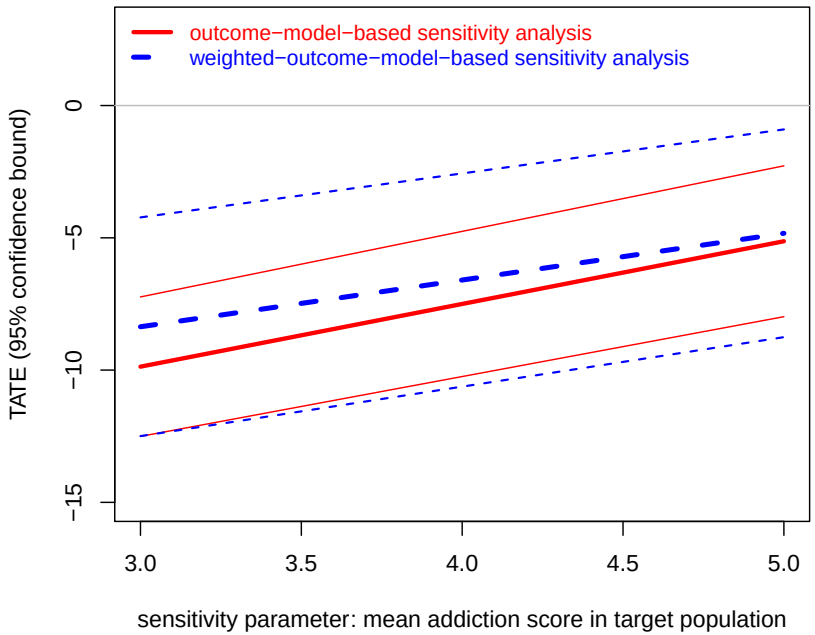
2. Weighted-outcome-model-based sensitivity analysis
   0. weight trial sample to resemble target population w.r.t Z, X
   i. obtain estimate for $\text{E}[Z \mid P = 1]$ and specify range for $\text{E}[V \mid P = 1]$
   ii. estimate $\beta_a, \beta_{za}, \beta_{va}$ using the weighted trial data
   iii. combine

Example of a V case

Smoking cessation intervention for heavy smokers among attendants of alcohol/substance abuse treatment: SATE = 10 fewer cigarettes per day

- Z: being African-American, baseline daily number of cigarettes
- V: baseline addiction score; $E[V \mid S = 1] = 4.05$

Target pop: people who seek alcohol/substance treatment who smoke heavily



What about a fully unobserved moderator (U)?

- What if the moderator is unobserved even in the trial?
- This harder, in part because of strong assumptions about the distribution of the covariates and U
- One approach in Nguyen et al. (2017), but not very general
- Hopefully this sort of scenario is not very common . . .

What about multiplicative treatment effects?

- Big limitation: the assumption of additive treatment effects
- Want flexibility in choosing effect scale
- Effects may be less heterogeneous on one scale than on another
- Current work extending these methods to binary and other types of outcomes

Outline

Everyone wants to assume that study results generalize

- But very few statistical methods exist
- At this point, lots of “hand waving," qualitative statements
- Need more statistical methods to quantify and improve external validity
    - For both study design and study analysis

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 in rigorous evaluations
- 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
  - These are coming along

Data a primary limiting factor

- Right now we have very little information on factors that influence effects or participation in trials
- Sometimes hard to find population data
- Trial data also often not publicly available
- Even harder to find population data that has the same measures as trial of interest
- Sensitivity analyses one approach for recognizing this limitation

Recommendations (1)

- 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

Recommendations (2)

- Standardize measures
  - At least make it more feasible to combine trial and population information
  - e.g., release individual items, not just scale scores, or at least scale them to the national population
  - e.g., consider use of common measures across studies
- 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
  - Also more work on methods such as response surface models to combine information from multiple sources
  - (Will also need more work on measure comparability and how to combine multiple measures)

Current work

- With Rob Olsen et al. (Westat): What about using existing trials to predict impacts for individual sites? How well can we do that (if at all)?
- With Issa Dahabreh (Brown): Developing doubly robust estimators for population effects, translating for PCORI audience
- With Trang Nguyen (Hopkins): Extensions of the sensitivity analyses to handle an unobserved effect moderator
- With Hwanhee Hong (Duke): Extending these methods to meta-analysis contexts with multiple trials
- With Ben Ackerman (fmr student; now Flatiron Health): Dealing with measurement error and generalizability, especially different measures in trial and population
- With Daniel Westreich (UNC): Translating these ideas to epidemiologists and the concept of “target validity"
- More on combining experimental and non-experimental evidence?
- ???

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)

For more information . . .

www.elizabethstuart.org

estuart@jhu.edu

@lizstuartdc

Funding thanks to NIH (K25MH083846), NSF (DRL-1335843), IES (R305D150003), PCORI (ME-1502-27794)

References, with thanks to all my co-authors

- Bell, S.H., Olsen, R.B., Orr, L.L., and Stuart, E.A. (2016). Estimates of external validity bias when impact evaluations select sites non-randomly. *Education Evaluation and Policy Analysis* 38(2): 318-335.
- 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.
- Dahabreh, I., Hernan, M., Robertson, S., Steingrimsson, J., and Stuart, E.A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*. 39:14:1999-2014.
- 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.
- Kern, H.L., Stuart, E.A., Hill, J., and Green, D.P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness* 9(1): 103-127.
- Mercer, A.W., Kreuter, F., Keeter, S., and Stuart, E.A. (2017). Theory and practice in nonprobability samples: Parallels between causal inference and survey inference. *Public Opinion Quarterly* 81: 250-279.
- Nguyen, T., Ebnesajjad, C., Cole, S.R., and Stuart, E.A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics* 11(1): 225-247.
- Orr, L.L., Olsen, R.B., Schmid, I., Shivji, A., Bell, S., and Stuart, E.A. (2019). Using the results from rigorous multi-site evaluations to inform local policy decisions. *Journal of Policy Analysis and Management*. 38:4: 978-1003.
- Stuart, E.A., Bell, S.H., Ebnesajjad, C., Olsen, R.B., and Orr, L.L. (2017). Characteristics of school districts that participate in rigorous national educational evaluations. *Journal of Research on Educational Effectiveness* 10(1): 168-206.
- 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., and Rhodes, A. (2017). Generalizing Treatment Effect Estimates from Sample to Population: A case study in the difficulties of finding sufficient data. *Evaluation Review* 41(4): 357-388.
- Westreich, D., Edwards, J., Lesko, C., Cole, S., and Stuart, E.A. (2019). Target validity and the hierarchy of study designs. *American Journal of Epidemiology* 188(2): 438-444.