Combining trial and population data to estimate population average treatment effects

Elizabeth Stuart
Executive Vice Dean, Professor

www.elizabethstuart.org

@lizstuartdc

July 13, 2022

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Outline

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 monoclonal antibodies 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

Generalization more broadly

- There are lots of reasons why results from randomized trials may not generalize
  - Scale-up problems, different contexts, different implementation
  - UTOSTi framework (Cronbach, Cook): Units, Treatments, Outcomes, Settings, Time
- Today will focus on differences due to differences between a trial sample and a population in characteristics that moderate treatment effects
- Provide statistical ways to think through some of the challenges of generalizability
- 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)
- Note 3: These ideas go back many years, back to Cook and Campbell in particular (threats to validity)

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 Δ_{SX}	Δ_{SU}	Treatment Selection Bias Δ_{TX}	Δ_{TU}	Total Bias
Ideal experiment	0	0	0	0	0
Typical experiment	Big	Big	0	0	Big
Typical non-exp study	Small	Small	Big	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

Half of H.I.V. Patients Are Women. Most Research Subjects Are Men.

Trials of vaccines and treatments have not included enough female participants. Now that scientists are exploring possible cures, the need to enroll women is greater than ever.

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

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

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Increasing evidence on differences between trial samples and populations

- Varma et al. (2021) showed older adults and Black patients underrepresented in trials of cancer therapeutics
- 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

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

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
- Note: Registries may be quite useful in the design stage!

Analysis possibilities for settings with multiple studies

- Meta-analysis
  - 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

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)
- 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., using BART; Kern et al., 2016)
- Doubly robust approaches that combine weighting and outcome models (Dahabreh et al., 2020)

Presumed data structure

	S (trial membership)	Y (outcome)	A (treatment)	X_1 (covariate 1)	X_2 (covariate 2)	...	X_p (covariate p)
Trial data set	1	✓	✓	✓	✓		✓
Population data set	0	×	×	✓	✓		✓

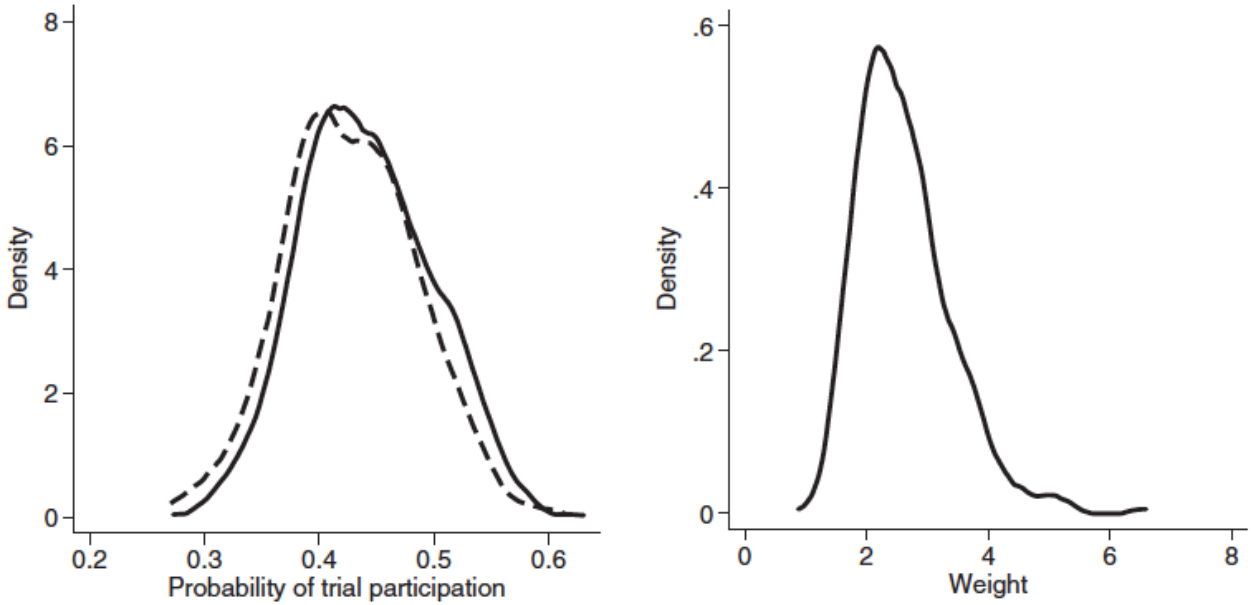
JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Example: Coronary Artery Surgery Study (CASS; Dahabreh et al., 2020)

- Randomized trial nested within a cohort study
- Trial compared surgery + medical therapy to medical therapy alone for patients with chronic coronary artery disease
- 780 patients randomized; 1319 declined
- Outcome: 10 year risk of death from any cause
- Covariates: age, disease severity, ejection fraction, other measures of coronary health
- Subjects in trial weighted by inverse odds of participating in the trial to weight to the

No big differences between trial participants and non-participants

FIGURE 1 Kernel densities for the estimated probability of trial participation for trial participants (left panel, solid line) and nonparticipants (left panel, dashed line), and the estimated weights for trial participants (right panel). The weights for trial participants are equal to the inverse of the estimated odds of trial participation times the inverse of the estimated probability of receiving the treatment actually received, as defined in Section 5.2



JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Individual covariate differences

	Non-randomized group		Randomized group	
Covariate	Surgery	Therapy	Surgery	Therapy
Age	51.3	50.6	51.4	50.9
Angina	84%	76%	77%	78%
History of MI	55%	60%	57%	63%
Ejection fraction	60.2	60.1	60.9	59.8

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Generalized effect estimates in CASS: Generally similar

Estimator	Risk difference
Trial only	-3.0 (-8.7, 2.7)
Outcome model	-1.3 (-7.9, 4.2)
Weighted (std. weights)	-1.9 (-7.2, 4.8)
Doubly robust	-1.4 (-7.3, 4.7)

The key assumptions

- Positivity: Everyone in the target population had a non-zero probability of participation in the trial
- Ignorability: All effect moderators related to sample selection observed and adjusted for
  - Rare to have extensive data overlap between sample and population
  - Need sensitivity analyses to assess robustness of results to violation of this assumption (Nguyen et al., 2017)
  - But well designed population data, and trial data aligned with those, can greatly help make this assumption plausible
  - Note: Having outcome data under "control" in the population can also help test this assumption (Hartman et al., 2015)

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)

If there is a partially unobserved effect modifier (V)

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

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

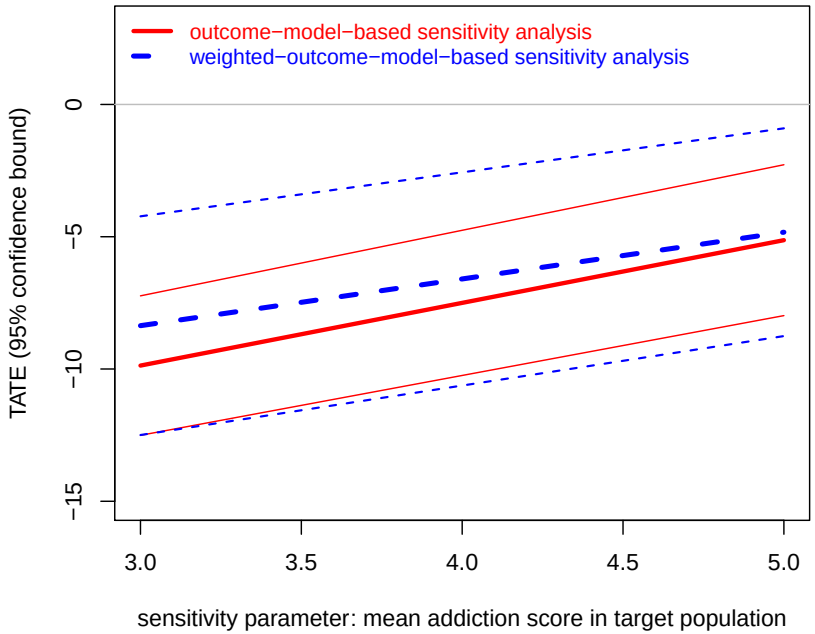
2. Weighted-outcome-model-based sensitivity analysis
   0. weight trial sample to resemble target population w.r.t Z, X
   i. obtain estimate for $\text{E}[Z \mid P = 1]$ and specify range for $\text{E}[V \mid P = 1]$
   ii. estimate $\beta_a, \beta_{za}, \beta_{va}$ using the weighted trial data
  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

Example

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 . . .

Outline

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

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 should researchers do?

- Generate more information on:
  - Factors that influence treatment effect heterogeneity
  - Factors that influence participation in rigorous evaluations
- Collect common measures in trial and population data
- Report how study samples were obtained
  - CONSORT and related diagrams a step towards this
- Methods that allow for the differences between trial and population on these factors
  - These are coming along
- (Remember too that external validity does not always need to be the goal! Depends on the stage of the research, context, etc.)

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Recent/current work

- With Rob Olsen et al.: What about using existing trials to predict impacts for individual sites? How well can we do that (if at all)?
- With Issa Dahabreh (Harvard): Developing doubly robust estimators for population effects, generalizing in multilevel contexts
- 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"
- With Hwanhee Hong, Trang Nguyen, Ben Goldstein (JHU/Duke): Combining small experiments and large-scale non-experimental data to estimate effect heterogeneity (NIMH and PCORI projects)

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), WT Grant Foundation

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.

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.

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.

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):438-443.

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH