Opportunities for, and complexities of, using real-world data

Elizabeth Stuart
www.biostat.jhsph.edu/∼estuart
estuart@jhu.edu

April 1, 2019

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Outline

1. Introduction
2. Non-experimental comparison group designs
3. Generalizing trial results to target populations
4. Comparative interrupted time series designs
5. Conclusions

Outline

(Some) potential uses for real world data

- Non-experimental comparison group designs
  - e.g., using historical control data with a single-armed trial
  - e.g., to study effectiveness in real world practice, post-approval
  - Propensity score/comparison group designs, instrumental variables
- Generalizing randomized trial results
  - e.g., to understand similarity of trial participants to populations of interest
  - e.g., to project impacts from trial into that population, perhaps by combining experimental and non-experimental evidence
  - New methods for generalizing/transporting effect estimates
- Policy evaluation/health services research
  - e.g., study different payment models for mental health care
  - e.g., study policy change in a hospital system
  - Interrupted time series type methods

What data am I thinking about here?

- Large administrative datasets
- Medical claims, electronic health records
- Registry data (including Scandinavian)
- Other system level administrative data

Outline

Non-experimental comparison group designs

- Main idea: compare people receiving and not receiving some intervention of interest
- Use “equating" methods (propensity scores, other matching methods) to make exposure groups look as similar to one another as possible
- Huge literature in this area
- You’ve already heard some about these methods already today

Propensity score methods as one approach

- Main problem in non-experimental studies is confounding: treatment and comparison individuals may be very different from one another on lots of factors
- Propensity scores commonly used as key design tool in such studies
- Goal is to replicate a randomized experiment as much as possible, by forming groups similar on the observed covariates
- (Relies on assumption that there are no unobserved differences between the groups once we condition on the observed ones; can also do sensitivity analysis regarding this assumption)

Effects of psychosocial therapy after suicide attempt

- Use Danish registry data (on population of people in Denmark) to compare outcomes of individuals who received psychosocial therapy after a suicide attempt to similar individuals who didn't
  - Very large sample, allows long follow-up, extensive covariates available
- Suicide prevention clinics began operation in Denmark in 1992, now nationwide
- Erlangsen et al. (2014)

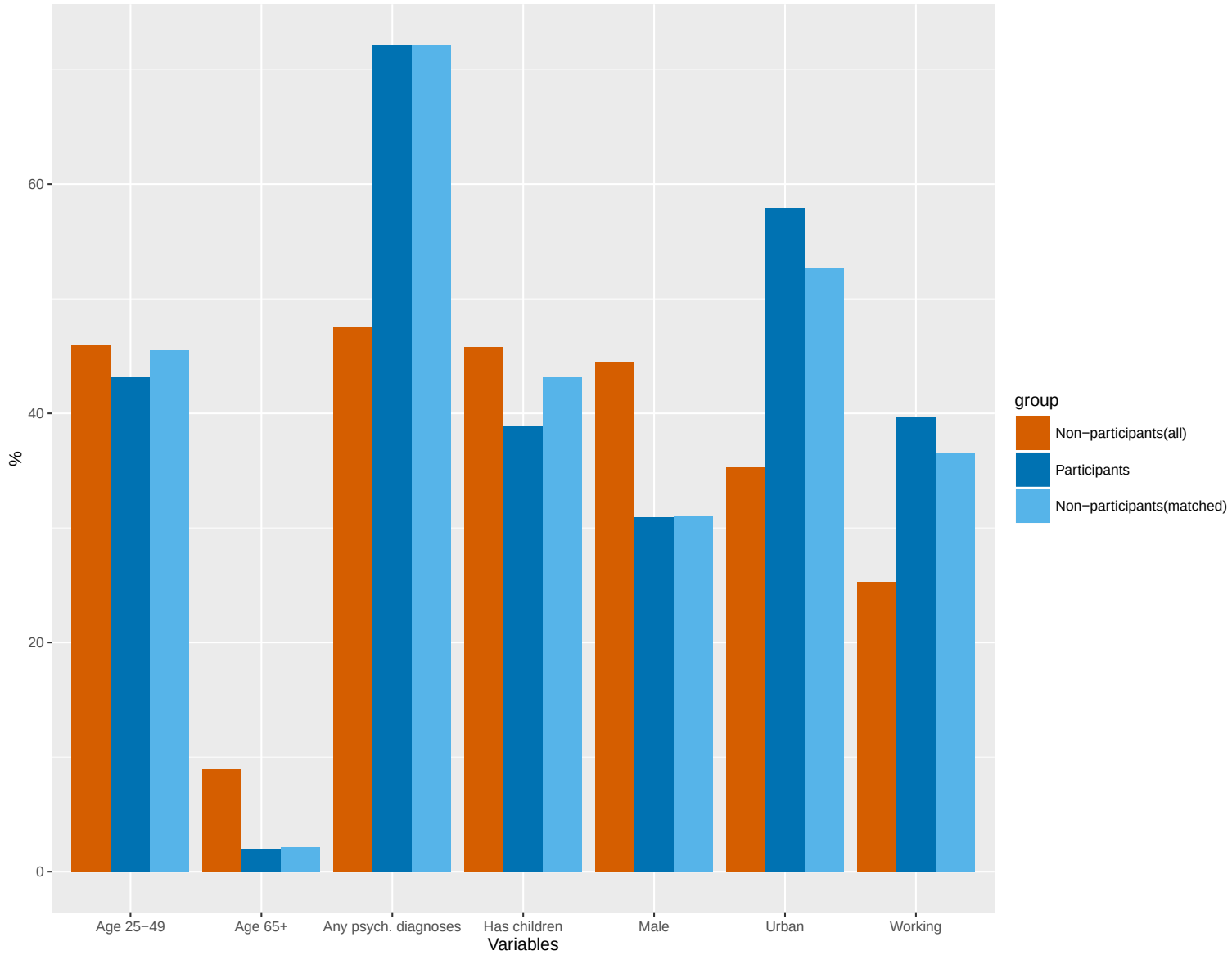
What do propensity scores do?

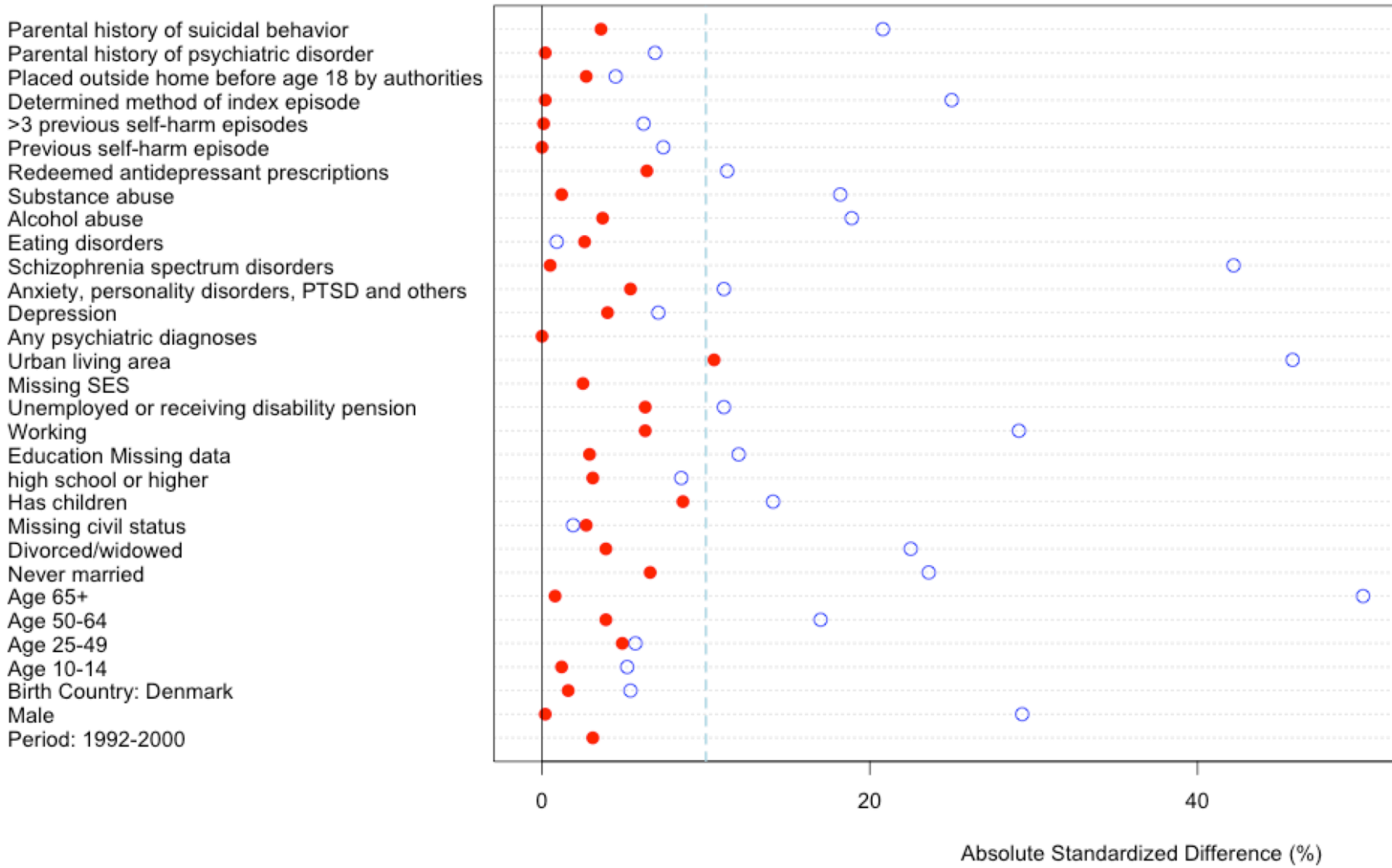
- The problem is that it is hard to find similar groups with respect to all covariates individually
- Propensity scores give a particular type of dimension reduction that allows matching on just the propensity score, not dealing with each covariate individually
- Propensity score methods attempt to replicate two features of randomized experiments
  - Create groups that look only randomly different from one another (at least on observed variables)
  - Don’t use outcome when setting up the design

Application

- Used large-scale Danish registry data, with extensive data on individuals, family structure, medical history, etc.
- Data for 30 years on 5,678 people who had gone to a suicide prevention center, and 58,821 who had attempted suicide but not then gone to one of the centers
- Used 31 covariates: demographics, previous suicide attempt, method of attempt, family history, psychiatric history
- For each user of the clinics, found 3 individuals with similar propensity scores (who also had the same values for “any psychiatric disorder" and “previous deliberate self-harm")

Subjects look similar after the matching





Program was effective at reducing 10 year risk of . . .

- Repeat suicide attempts: OR=0.82; CI (0.75, 0.89)
- Death by suicide: OR=0.71; CI (0.56, 0.91)
- Death from any cause: OR=0.65; CI (0.57, 0.74)

- Findings robust to a potential unobserved confounder (Liu, Kuramoto, and Stuart, 2013)

Conclusions re propensity score methods

- Big picture idea: careful design of non-experimental studies (trial “emulation")
  - Clear statement of exposures of interest, outcome measure, timing, covariates, etc.
- Propensity score methods can be a useful tool for improving estimation of causal effects in non-experimental settings
- Can be used in large-scale population-based datasets
- Conclusions can then be drawn about the effects of exposures or interventions in that population
- But non-experimental studies will still have potential concerns about unobserved confounders that may bias the results

Outline

Generalizing trial results to target populations

- Concern with randomized trials is that they often enroll non-representative samples of people
- Unclear how well the results would translate to target populations of interest
- Can use real world data to help understand that
  - Use trial to predict outcomes in population, using covariate data in population
  - Combine trial results and non-experimental results from population (“cross design synthesis")

Quick example: ACTG trial (Cole & Stuart, 2010)

- Trial examined highly active antiretroviral (HAART) therapy for HIV compared to standard combination therapy
  - Intent-to-treat analysis: Hazard ratio of 0.51 (95% CI: 0.33, 0.77)
- Question: What would the effects of the treatment be if implemented nationwide?
- Stack trial data with data representing US population of newly infected individuals
- Trial and US population differ on variables that also moderate effects (age, race, sex)

Approach: 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
- (Outcome prediction model approaches exist too, e.g., BART, TMLE)
- Key assumption: No unobserved effect moderators that differ between sample and population

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

Conclusions

- Possible to use real world data to help understand how results from trials would translate
- Will depend on having consistent measurement of covariates across trial and population
- Note; These methods relevant when interested in population average treatment effect (e.g., for cost/benefit calculations), not for individual treatment decision making

Outline

Final approach: Comparative interrupted time series

- In cases of abrupt policy changes comparative interrupted time series methods can be used to estimate effects of that policy change
- Essentially compare pre and post periods
- Best if there are comparison sites without the change, to model trends over time
- Huge and growing literature, especially lately
- Lots of use of large-scale administrative data for these
- Example: Effects of federal mental health parity law on mental health service use (Stuart et al., 2017)
- Example: Gun control laws (Rudolph, Stuart, Vernick, and Webster, 2015)

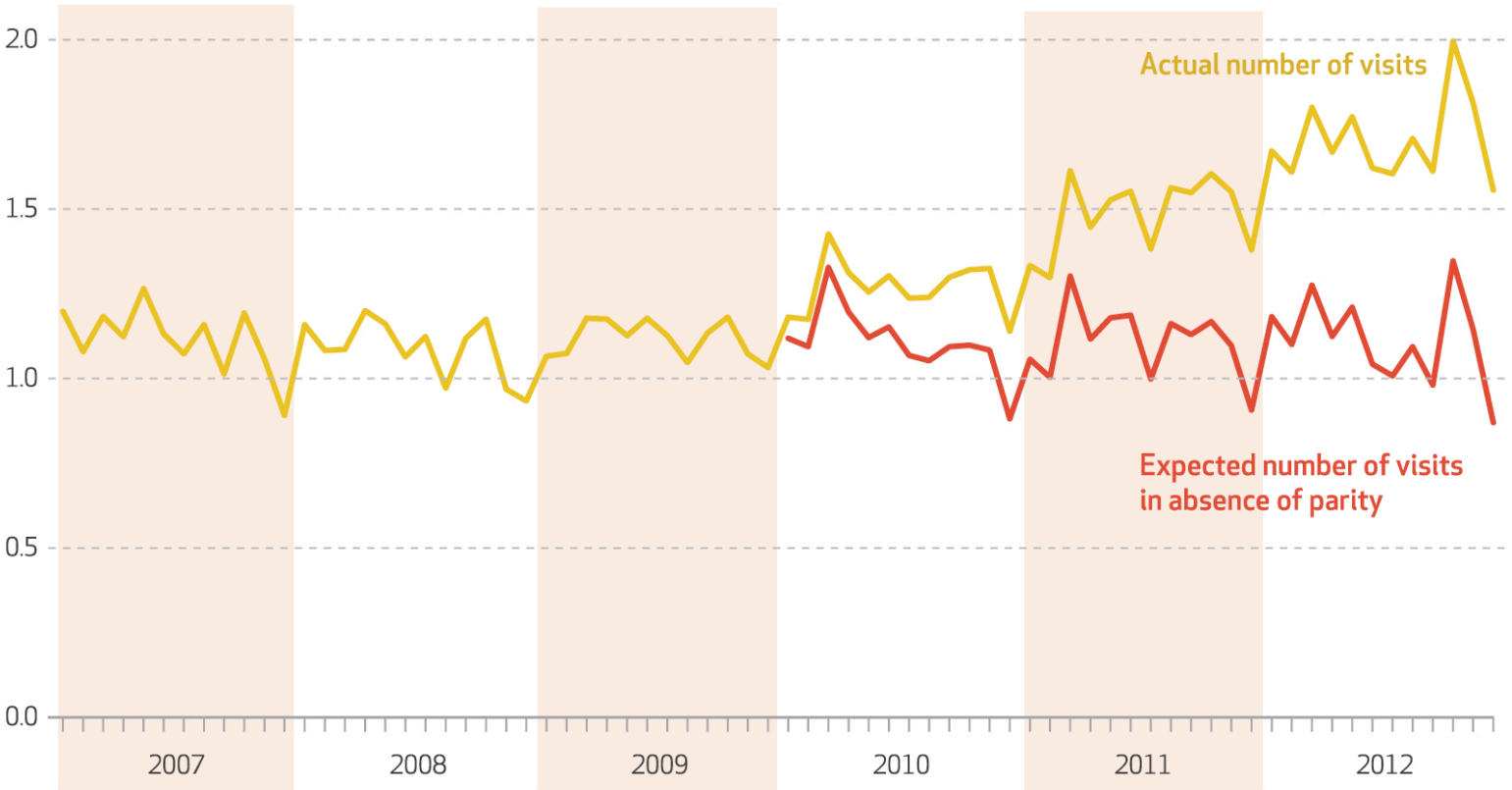
Effects of mental health parity law (Stuart et al., 2017)

- Used Marketscan claims data to look at effects of federal mental health parity law on kids with autism spectrum disorder
    - N=38,928
- Interrupted time series model: look at trends over time in service utilization, out of pocket spending
- Lots of complications: how to define the sample longitudinally, how to deal with confounding, what time scale to model, etc.
- Use ARIMA model to model trends over time and project out what would have happened in the absence of the intervention
- (Note: Unfortunately no comparison group here . . .)

Effects on service utilization (Stuart et al., 2017; Figure 3)

EXHIBIT 3

Associations between the federal parity law and outpatient mental health and functional therapy visits, among youth using those services



SOURCE Authors' analysis of data from the Truven Health MarketScan Research Database, 2007–12. **NOTE** Functional therapy visits include speech and language therapy and occupational and physical therapy.

Outline

What are some of the challenges in using this data in these ways?

- Measurement timing sometimes messy
  - e.g., unclear index date for comparison group members
  - Challenges in EHR because people only show up in the data when they go to the doctor
- Measurement error in covariates, exposures and outcomes
  - Need to think about implications for each study; use of validation samples?
  - And if combining data (e.g., trial and population), challenges if variables measured differently in different sources
- Missing confounders/moderators
- (Honestly the epidemiologists have thought a lot more about these issues than have statisticians!)

Conclusions

- Lots of opportunities, and plenty of challenges too
- Huge literature on these topics
- Let's make sure statisticians are at the table when thinking about how to use (and not use) the data!

For more information . . .

- http://www.biostat.jhsph.edu/∼estuart/
- estuart@jhu.edu
- Funding thanks to NIH, NSF, IES, PCORI

Selected references

- General
  - Davy, R. . . . Stuart, E.A. (2018). Exploring whether a synthetic control arm can be derived from historical clinical trials that match baseline characteristics and overall survival outcome of a randomized controlled trial: Case study in non-small cell lung cancer. Friends of Cancer Research white paper. https://www.focr.org/sites/default/files/SCA%20White%20Paper.pdf
  - Franklin, J.M., and Schneeweiss, S. (2019). When and how can real world data analyses substitute for randomized controlled trials? *Clinical Pharmacology & Therapeutics*.
  - Stoto, M., Oakes, M., Stuart, E., Priest, E.L., Savitz, L. (2017). Analytical Methods for a Learning Health System: 2. Design of Observational Studies. eGEMs (Generating Evidence & Methods to improve patient outcomes). 5(1): 29.

- Comparison group methods
  - Erlangsen, A., . . . , Stuart, E.A., et al. (2015). Short and long term effects of psychosocial therapy for people after deliberate self-harm: a register-based, nationwide multicentre study using propensity score matching. Lancet Psychiatry 2(1): 49-58
  - Jackson, J.J., Schmid, I., and Stuart, E.A. (2017). Propensity scores in pharmacoepidemiology: Beyond the horizon. Current Epidemiology Reports. Topical collection on pharmacoepidemiology. Published online 6 November 2017.
  - Stuart, E.A. (2010). Matching Methods for Causal Inference: A review and a look forward. Statistical Science 25(1): 1-21

- Generalizing treatment effects
  - 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., Robertson, S., Stuart, E.A., Hernan, M., and Tchetgen Tchetgen, E. (in press). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Forthcoming in Biometrics.
  - 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.*
  - 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.

- Interrupted time series
  - https://diff.healthpolicydatascience.org/
  - Rudolph, K., Stuart, E.A., Vernick, J., and Webster, D. (2015). Association between Connecticut's Permit-to-Purchase Handgun Law and Homicides. American Journal of Public Health 105(8): e49-54.
  - Stuart, E.A., McGinty, E.E., Kalb, L., Huskamp, H.A., Busch, S.H., Gibson, T.B., Goldman, H., and Barry, C.L. (2017). Increased Service Use Among Children With Autism Spectrum Disorder Associated With Mental Health Parity Law. Health Affairs (Millwood) 36(2): 337-345. PMID: 28167724