Propensity score methods in the context of covariate measurement error

Elizabeth Stuart
Associate Dean for Education, Professor
www.biostat.jhsph.edu/∼estuart

16 November 2017

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Outline

1. Introduction
2. Methods for handling covariate measurement error in propensity score methods
3. Handling measurement error through multiple imputation
4. Simulation
5. Application
6. Teasers of other approaches
7. Conclusions

Outline

Big picture

- Goal: Estimate the causal effect of receiving one treatment relative to a comparison condition
- Non-experimental studies use naturally occurring groups of individuals, some who got the treatment and some who got the comparison condition
- Problem is potential “selection bias":
  - Individuals in treatment group may differ quite a bit from those in the comparison group
  - Thus, differences in outcomes may be due to those baseline differences, not to the treatment itself
- Many approaches try to limit selection bias by adjusting for (or matching on) covariates before estimating effects
- But what if those covariates are not measured perfectly?
  - e.g., self-reported measures of height or weight, imperfect measures of blood pressure, latent constructs for depression or disability

More formal: Potential outcomes model for defining treatment effects

- Y(0)=potential outcome under comparison condition
- Y(1)=potential outcome under treatment condition
- T=treatment variable (1=treatment, 0=control)
- We observe $Y_{\text{obs}} = T * Y(1) + (1 - T) * Y(0)$
  - The “fundamental problem of causal inference"
- The treatment effect for individual is D=Y(1)-Y(0)
- Interest usually in average treatment effects across a population: E(D)=E(Y(1))-E(Y(0))
- Goal in a non-experimental study: Use treatment group to estimate E(Y(1)) and the comparison group to estimate E(Y(0)), but accounting for the fact that the treatment and comparison groups are not necessarily random samples from the population of interest

Propensity score methods

- Propensity scores provide a way of “equating" the groups to make the treated and comparison groups look as similar as possible on the observed covariates
  - Propensity score = predicted probability of receiving the treatment, given observed covariates
- Typical ways of using propensity scores: matching, weighting, subclassification (Stuart, 2010)
- Today will focus on Inverse Probability of Treatment Weighting (IPTW)
  - Treated group weights: 1/p
  - Comparison group weights: 1/(1-p)
- Separation of “design" and “analysis": Outcomes not (typically) used in the propensity score process

The standard assumption underlying propensity score analyses

- Most propensity score analyses rely on assumption of unconfounded treatment assignment:
  - $T \perp (Y(0), Y(1))|X$
  - No unobserved differences between treatment and control groups, given the observed covariates X
- What if treatment assignment actually depends on true X but all we observe is a mis-measured version of it, W?
  - e.g., decision to take a new treatment depends on true underlying health status, but all we have are proxies for it
  - e.g., decision to take a new treatment depends on blood sugar levels, but all we have are claims data
- Steiner et al. (2011) and others have shown that bias reducing ability of propensity scores can be diminished due to covariate measurement error

Outline

Some potential solutions

- Latent variable approach: Model true, underlying latent variable (Raykov, 2012)
  - Investigated in context of regression adjustment for propensity scores
  - Requires multiple indicators for the true covariate
  - (With Trang Nguyen I am investigating extensions of this approach; initial results suggest best approach is to estimate a “full" factor model that includes T in the factor model and then include the factor score in the propensity score model)
- Corrected propensity score weighting strategy (McCaffrey et al., 2011; McCaffrey and Lockwood, 2016)
  - For propensity score weighting only
  - Assumes classical measurement error
  - Requires some external calibration information

- Empirical expressions for resulting bias (Ogburn and VanderWeele, 2013)
  - Under certain assumptions, show that controlling for a mismeasured covariate will result in estimate between the crude and true effect measures
  - Can help bound the effect
- Plus 3 other approaches I will briefly mention (SIM-EX, Bayesian model, and sensitivity analysis) and another I will cover in depth (multiple imputation)

Outline

The multiple imputation approach

- Main idea: Use a source of information on the relationship between W and X to multiply impute values of $X|W$
  - Intuitively, should account for uncertainty in imputations of X
- For now, will assume that we have some external validation sample with data on X and W (and possibly other common variables Z)
  - Things more complex without this
  - (And are easier if internal validation data available)
- Imputations actually nested: m values of parameters drawn, then n imputations from each parameter draw
- Run propensity score approach within each imputed dataset
- Combine effect estimates across imputed datasets
- Has appeal due to flexibility (as with normal MI)

But . . . the simple approach doesn't work

- Can't just generate model of X given W in the calibration sample and then apply that in the main sample to predict X
- Model uncongeniality if imputation model doesn't incorporate T and Y

Multiple imputation - external calibration (MI-EC)

- Instead use MI-EC, which uses joint distribution of all variables to generate imputations of X (Guo, Little, and McConnell, 2012)
  - Constructs posterior distribution of $f(X|T, Y, Z, W)$
- Gets information on joint distribution of X and W from the validation sample
- Gets information on joint distribution of W, T, Y, Z from the main sample
- Key assumptions:
  - Multivariate normality:
    - $f(Y, T, Z, X|W) \sim N(\beta W, \Sigma)$
  - Strong version of non-differential measurement error
    - $f(Y, T, Z|X, W) = f(Y, T, Z|X)$
    - Measurement error can not depend on Z, T
    - Standard assumption would have Z, T as conditioned on

Specific steps for using MI-EC in the context of propensity score analysis

1. Generate multiple (nested) imputations of the true covariate X using MI-EC
2. For each imputation:
    1. Estimate propensity scores
    2. Use a propensity score approach to estimate treatment effects (we will use weighting)
3. Combine results across nested multiple imputations, using standard MI combining results for nested imputations

Outline

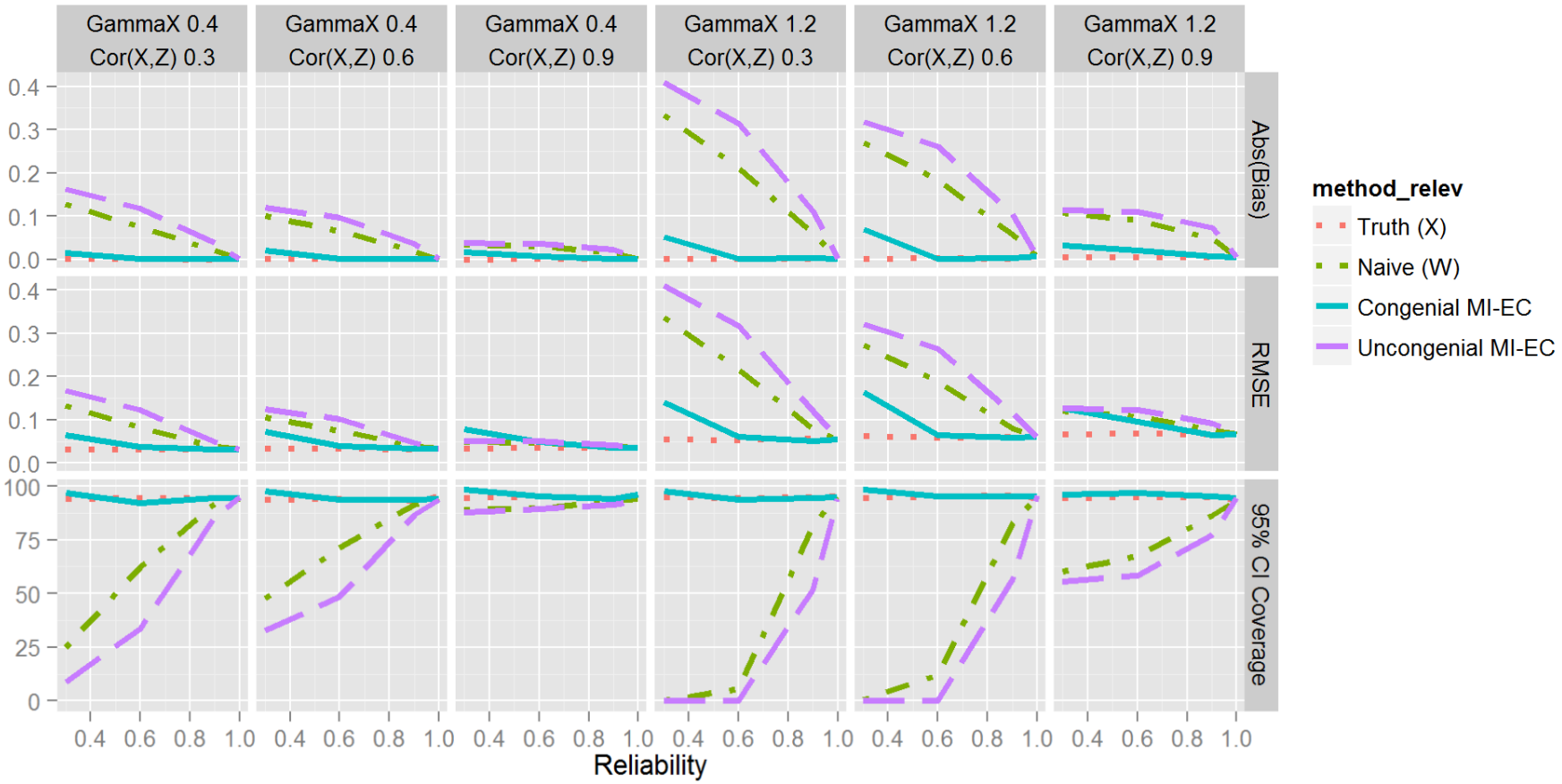
Simulation set-up

- X,Z jointly normally distributed, means 0, correlation ρ
- Treatment a logistic function of X, Z
- Measurement error model: $W|X \sim N(X, \sigma^2)$
- Y a function of T, X, and Z:
  $Y|T, X, Z \sim N(\Delta T + \delta_X X + \delta_Z Z, \tau^2)$
- (So X the true confounder, but we only observe a mis-measured version of it, W)
- Assume $N_{\text{main}} = 2500$, $N_{\text{val}} = 500$
- Varied parameters, especially correlation between X and W

Methods compared

- Naive (just using W)
- Gold standard (using X)
- MIEC using just Z, W (uncongenial)
- MIEC using W, Y, Z, and T (congenial)

Simulation results



Summary of simulation results

- Ignoring the measurement error leads to bias
- More bias if X strongly related to treatment assignment
- Less bias if X and Z strongly correlated
- Less bias if X and W strongly related (high reliability)
- Even if *Z* not predictive of Y, including it in the procedure helps a lot (“auxiliary variable")
- Using MI-EC can correct for most of the bias
- But using an uncongenial MI-EC (with only Z) worse than naive approach (this is like a naive imputation using the validation sample to fit the imputation model)

Outline

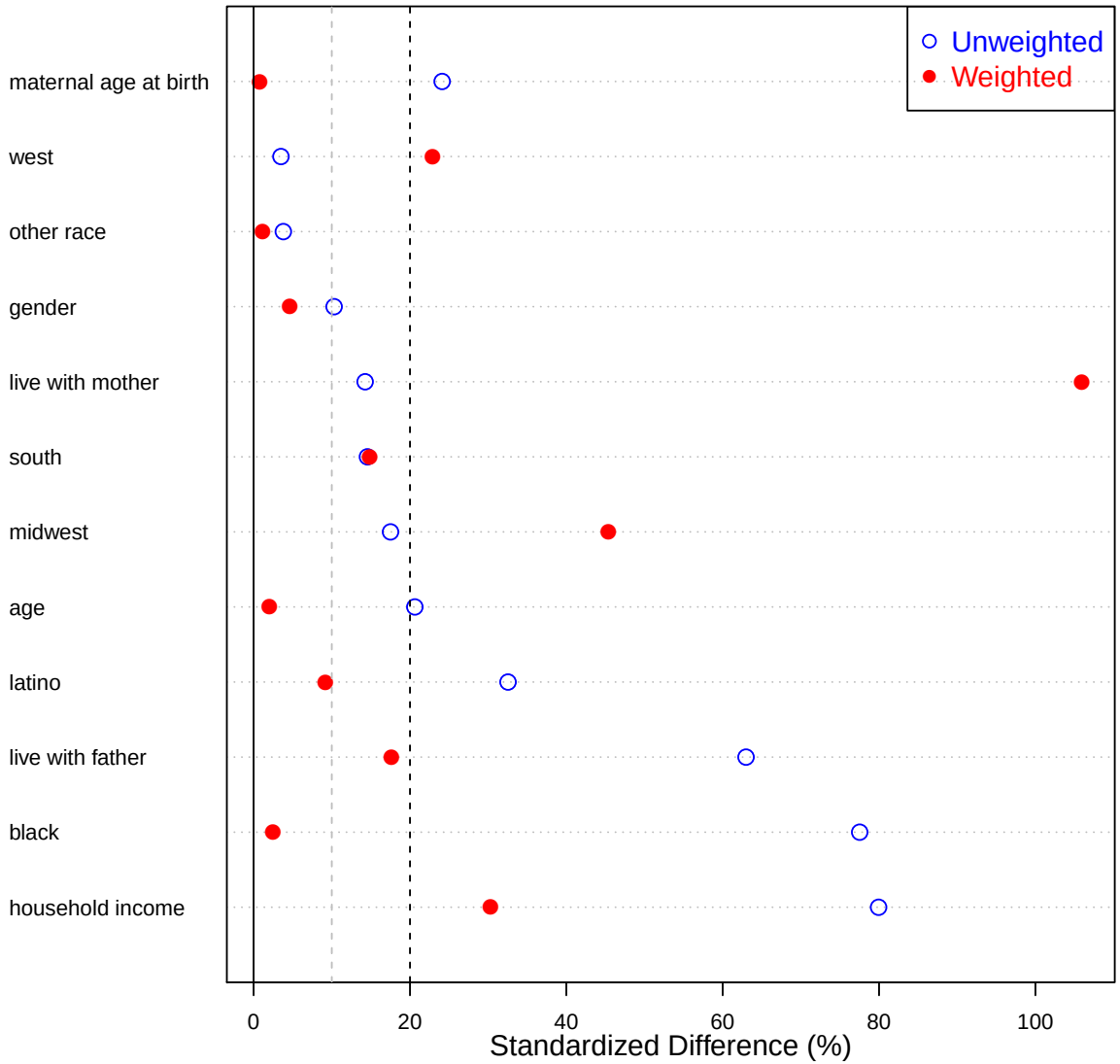
Living in a disadvantaged neighborhood and mental health outcomes

- Interest in the consequences of living in a disadvantaged neighborhood on a variety of outcomes, including mental health and substance abuse
- Data: National Comorbidity Survey Replication Adolescent Supplement (NCS-A)
- Nationally representative survey of approximately 10,000 adolescents
- Established score for neighborhood disadvantage used: lowest tertile considered the “treatment" group
- Compare adolescents in lowest tertile with those in upper tertiles

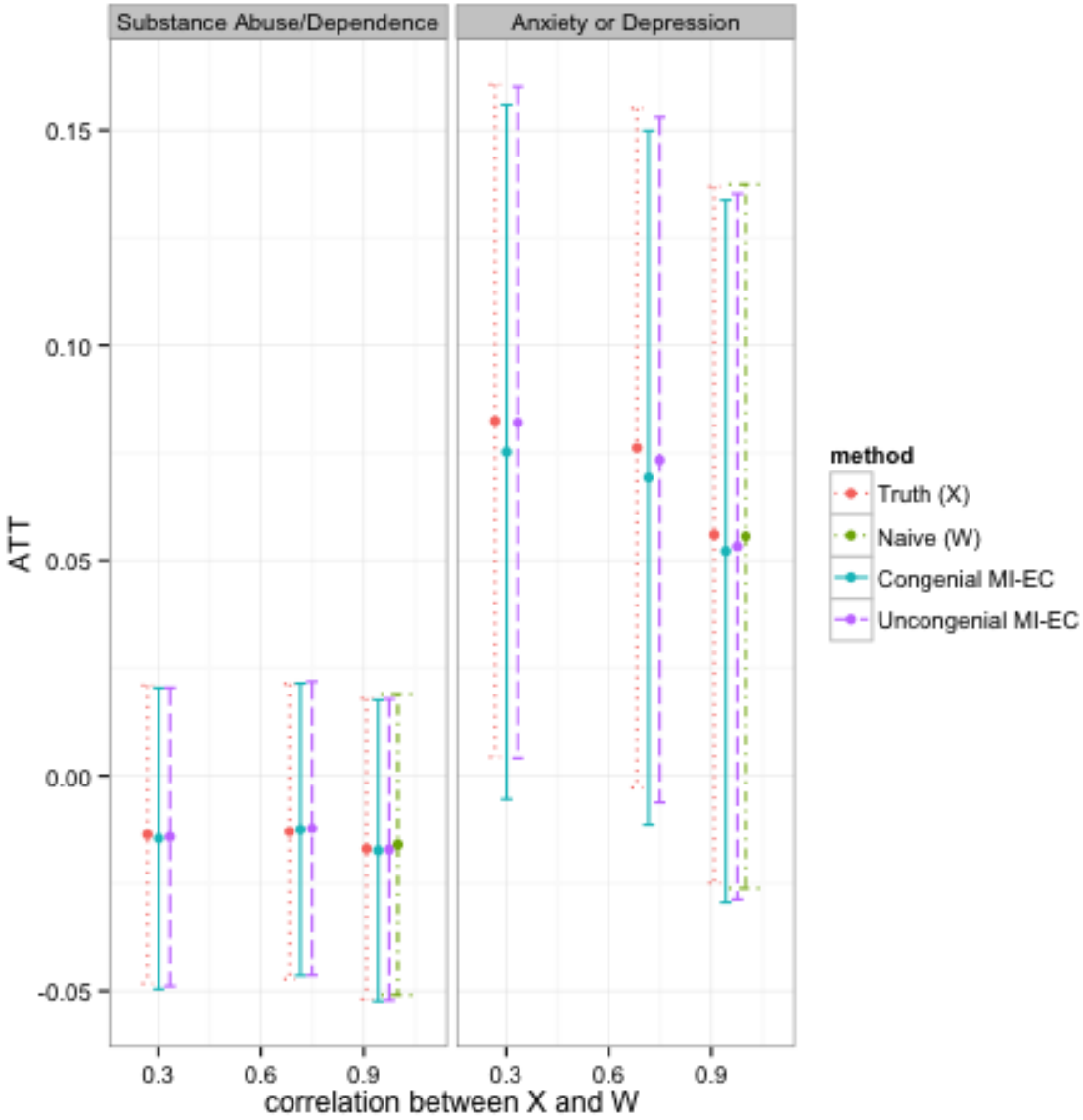
Details of application

- Covariates available: Gender, age, race/ethnicity, family income, family structure, mother's age at birth
- True covariate: Mother's report of her age at birth of child (not always available)
- Covariate measured with error: Child's report of maternal age at birth
  - In reality, not much measurement error ($\rho = .94$)
  - So have 2 additional scenarios where we add on additional random noise to W ($\rho = .72$, $\rho = .3$)
- (Actually restrict the sample to those with both versions available, to use as a check)
- Outcomes: Past-year substance abuse or dependence, past year depression or anxiety
- Use a random subset of 400 adolescents as the validation sample

Covariate balance



Outcome results



Conclusions from application

- Not much difference across methods
- Amount of measurement error also doesn't seem to matter much
  - Maybe because lots of other covariates being used?
- Should treat these as illustrative, not as definitive substantive conclusions
  - Using a subset of the data (those with maternal and adolescent report)
  - Complex survey design not incorporated into analysis

Outline

Simulation-Extrapolation (SIM-EX); Lenis et al. (2016)

- Involves adding additional measurement error to the data, estimating effects given increasing amounts of measurement error, and then extrapolating back to 0 measurement error
- Explored a mean-reverting measurement error structure: $W_i = X_i + \tau_1[X_i - E(X_i)] + \sigma\epsilon_i$
- Examine asymptotics of doubly robust treatment effect estimator that uses SIM-EX to adjust for the measurement error
- Simulations (inc. based on real data) show good performance
- Also see McCaffrey and Lockwood (2014) for classical measurement error case

Bayesian model (Hong, Rudolph, and Stuart, 2017)

- More complicated measurement error structures may involve differential measurement error
- e.g., measurement error depends on another variables (inc. possibly the treatment indicator!)
- Develop a Bayesian model with parameters that allow for differential measurement error (both location and/or scale)
  - $W|X \sim N(X + \gamma A, \sigma^2_{w|x,a=0}(1 + \delta A)^2)$
- Simulations (and intuition) show this measurement error structure can be particularly problematic!

- A particular complication is that there is often limited data on the measurement error parameters; may involve non-identified parameters
- Consider two approaches:
  - Joint Bayesian model estimating propensity score and outcome models together
  - Two step approach that first generates posterior draws of X, and then uses those in outcome model
- Find that bias can be quite large if differential location across groups
- Heteroskedasticity doesn't matter as much
- Prior can matter a lot; need to specify carefully
- If X is a weak predictor of outcome, naive approach fine
- If X is a strong predictor of outcome, joint Bayesian approach best (although potentially controversial)

Sensitivity analysis approach (Rudolph and Stuart, in press)

- Can treat the measurement error explicitly as an unobserved confounder, use strategies for unobserved confounding
- Examine use of established approaches for unobserved confounding in non-experimental studies, adapt for measurement error
- Examines classical and differential measurement error
- Find good performance of bias formulas (VanderWeele and Arah) or a version of propensity score calibration that uses weighted least squares
    - (Standard propensity score calibration doesn't work well because of strong assumption)
- Standard Rosenbaum sensitivity analysis approach does not work well here; hard to interpret the needed parameters and only appropriate for matching

Outline

Further directions

- More investigation of when measurement error matters
  - May not be a lot of problem if classical error, and not a super strong confounder
  - But differential measurement error can cause a lot of problems
- Comparison of methods
- Further investigation of consequences of model misspecification or violation of assumptions
- What if validation sample not available?
- What if the validation sample is not representative of the main sample; adjust for that?
- Does the propensity score approach itself matter?

Conclusions

- Measurement error common and a potentially important concern in propensity score methods
- MI-EC and other strategies can be an effective strategy for handling measurement error in the context of propensity score analyses
- One limitation of some of them is need to include Y in the imputation procedure; may violate the clean separation of “design" and “analysis" that is one of the key benefits of propensity score methods
- Many more questions to be answered!

References

- Thanks to NIH for funding this work (NIMH 1R01MH099010)

- estuart@jhu.edu

- http://www.biostat.jhsph.edu/∼estuart/propensityscoresoftware.html
- Online 1 credit short course on propensity scores in JHSPH summer institute: http://www.jhsph.edu/departments/mental-health/summer-institute/courses.html

- Hong, H., Rudolph, K.E., and Stuart, E.A. (2016). Bayesian approach for addressing differential covariate measurement error in propensity score methods. *Psychometrika.* Published online October 13, 2016.
- 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.
- Lenis, D., Ebnesajjad, C.E., and Stuart, E.A. (2017). A doubly robust estimator for the average treatment effect in the context of a mean-reverting measurement error. *Biostatistics* 18(2): 325-337.
- Rudolph, K., and Stuart, E.A. (in press). Using sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods. Forthcoming in *American Journal of Epidemiology.*
- Stuart, E.A. (2010). Matching Methods for Causal Inference: A review and a look forward. *Statistical Science* 25(1): 1-21
- Webb-Vargas, Y., Rudolph, K.E., Lenis, D., Murakami, P., and Stuart, E.A. (2015). Applying multiple imputation for external calibration to propensity score analysis. *Statistical methods in medical research*. . Published online June 2, 2015.