Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects

Trang Quynh Nguyen[*] (tnguye28@jhu.edu), Cyrus Ebnesajjad,[*]
Stephen R. Cole[**] and Elizabeth A. Stuart[*]

[*]Johns Hopkins Bloomberg School of Public Health, [**]University of North Carolina
Funding thanks to NIDA T32-DA007292, NSF DRL-1335843

ENAR invited session 112:
Generalizing Clinical Data Across Studies/Populations
Austin TX, March 9, 2016

Background

- In public health/public policy, there are times we want to know the broad population effects of an intervention (e.g., a substance abuse treatment model) or policy change (e.g., handgun control laws)
- A randomized trial may have been conducted and an intervention effect estimated in the trial (SATE)
- SATE is different from the effect of the intervention if applied to a target population (TATE) if
  - there is intervention effect heterogeneity, and
  - the trial sample is different from the target population with respect to the distribution of factors that modify intervention effects
- Methods exist to estimate TATE, which require target population covariates data, especially data on effect modifiers
  - re-weighting trial sample to target population (Cole & Stuart, 2010)
  - model outcome in trial and predict outcome in target population (Kern et al., 2016)

The problem

But what if

- there is an effect modifier observed in the trial but we don't have data on it from the target population?
- we are concerned there might be effect modification that is not even observed in the trial?

Sensitivity analyses are needed.

Our purpose: develop simple procedures for use by substantive scientists.

Notation

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

Two datasets: trial data and a dataset representing the target population

S: sample membership (1=trial, 0=target population)

SATE $= \mathrm{E}_{S=1}[Y^1 - Y^0]$ and TATE $= \mathrm{E}_{S=0}[Y^1 - Y^0]$

X: non-effect-modifying covariates

Z: effect modifiers, observed in both samples

either V: effect modifier, observed in the trial but not the target population
or U: effect modifier, not observed in both samples

X, Z, V, U may be associated with S.

Toy example: A smoking reduction intervention

V case:	Randomized trial sample			Target population
	Treatment (n=200)	Control (n=200)	Full sample	sample (n=10,000)
Covariates				
X = Years of education: mean (SD)	12.06 (1.64)	12.11 (1.58)	12.08 (1.61)	11.02 (1.52)
Z = Gender: percent female	49.50	50.50	50.00	19.86
V = Years smoked: mean (SD)	7.36 (2.57)	7.50 (2.45)	7.43 (2.51)	not observed
Outcome				
Y = Cigarettes per week: mean (SD)	97.42 (6.00)	101.80 (5.29)	99.61 (6.06)	not observed

U case:	Randomized trial sample			Target population
	Treatment (n=200)	Control (n=200)	Full sample	sample (n=10,000)
Covariates				
X = Years of education: mean (SD)	12.06 (1.64)	12.11 (1.58)	12.08 (1.61)	11.02 (1.52)
Z = Gender: percent female	49.50	50.50	50.00	19.86
U ?				
Outcome				
Y = Cigarettes per week: mean (SD)	97.42 (6.00)	101.80 (5.29)	99.61 (6.06)	

Proposed sensitivity analyses

V case

- bias-formula-based method
- weighting-based method
- hybrid method

U case

- bias-formula-based method
- hybrid method

V case: V observed in trial but not in target population

Assumptions

- Sample ignorability for treatment effects: $(Y^1 - Y^0) \perp S|Z, V$
  - no other effect modifiers
    (or if any, they are independent of S conditional on Z, V)
- Overlap: the ranges of the effect modifiers in the target population are covered by their ranges in the trial
- Bias-formula-based and hybrid methods: an additive model for the potential outcomes of the form
  $\mathrm{E}[Y_i^t] = \beta_0 + \beta_t t + \beta_{zt} Z_i t + \beta_{vt} V_i t + f_{xzv}(X_i, Z_i, V_i)$
  - no three-way interaction ZVt
- Weighting-based method: distribution assumptions for V

V case: V observed in trial but not in target population

$$
\begin{aligned}
\mathrm{E}[Y_i^1] - \mathrm{E}[Y_i^0] &= \beta_t + \beta_{zt} Z_i + \beta_{vt} V_i \\
\Rightarrow \mathrm{SATE} &= \beta_t + \beta_{zt} \mathrm{E}_{S=1}[Z] + \beta_{vt} \mathrm{E}_{S=1}[V] \\
\mathrm{TATE} &= \beta_t + \beta_{zt} \mathrm{E}_{S=0}[Z] + \beta_{vt} \mathrm{E}_{S=0}[V] \\
\mathrm{SATE} - \mathrm{TATE} &= \beta_{zt}(\mathrm{E}_{S=1}[Z] - \mathrm{E}_{S=0}[Z]) + \beta_{vt}(\mathrm{E}_{S=1}[V] - \mathrm{E}_{S=0}[V])
\end{aligned}
$$

Bias-formula-based sensitivity analysis:

- Estimate SATE, $\mathrm{E}_{S=1}[Z]$, $\mathrm{E}_{S=1}[V]$, $\mathrm{E}_{S=0}[Z]$
- Estimate β_{zt}, β_{vt} using regression analysis of trial data
  $Y = \beta_0 + \beta_t T + \beta_{zt} ZT + \beta_{vt} VT + f_{xzv}(X, Z, V) + \epsilon$
- Specify a plausible range for $\mathrm{E}_{S=0}[V]$
- Get a range for the point estimate of TATE
  $\widehat{\mathrm{TATE}} = \widehat{\mathrm{SATE}} - \hat{\beta}_{zt}(\hat{\mathrm{E}}_{S=1}[Z] - \hat{\mathrm{E}}_{S=0}[Z]) - \hat{\beta}_{vt}(\hat{\mathrm{E}}_{S=1}[V] - \mathrm{E}_{S=0}[V])$

V case: V observed in trial but not in target population

If V were observed in both samples, could weight trial sample to resemble target population w.r.t the distribution of Z, V, and estimate TATE.

- The weights, $W_i = \dfrac{\mathrm{P}(S=0|Z_i, V_i)}{\mathrm{P}(S=1|Z_i, V_i)}$, are based on a model fit to the stacked dataset (combining the two samples)
- Rewrite $W_i = \dfrac{\mathrm{P}(S=0|Z_i)}{\mathrm{P}(S=1|Z_i)} \cdot \dfrac{\mathrm{P}(V=V_i|Z_i, S=0)}{\mathrm{P}(V=V_i|Z_i, S=1)}$

Weighting-based sensitivity analysis:

- Instead, obtain $\dfrac{\mathrm{P}(S=0|Z_i)}{\mathrm{P}(S=1|Z_i)}$ for trial participants
- Estimate $\mathrm{P}(V=V_i|Z_i, S=1)$ for trial participants
- Specify a plausible range for $\mathrm{P}(V|Z, S=0)$
- For each instance in the range, compute $\mathrm{P}(V=V_i|Z_i, S=0)$, assemble W_i, weight trial sample, and estimate TATE
- This gives a range for TATE with confidence limits

V case: V observed in trial but not in target population

If we just weight the trial sample using $W_i^{|Z_i} = \frac{P(S=0|Z_i)}{P(S=1|Z_i)}$ and estimate an ATE, we get a Z-adjusted ATE (zATE)

$$\begin{aligned} \text{zATE} &= \beta_t + \beta_{zt} E_{S=1,W^{|Z}}[Z] + \beta_{vt} E_{S=1,W^{|Z}}[V] \\ &= \beta_t + \beta_{zt} E_{S=0}[Z] + \beta_{vt} E_{S=1,W^{|Z}}[V] \\ \text{zATE} - \text{TATE} &= \beta_{vt}(E_{S=1,W^{|Z}}[V] - E_{S=0}[V]) \end{aligned}$$

Hybrid method sensitivity analysis:

- Weight trial sample using $W_i^{|Z_i}$ and estimate zATE, $E_{S=1,W^{|Z}}[V]$
- Estimate β_{vt} using regression analysis of unweighted trial data $Y = \beta_0 + \beta_t T + \beta_{zt} ZT + \beta_{vt} VT + f_{xzv}(X, Z, V) + \epsilon$
- Specify a plausible range for $E_{S=0}[V]$
- Get a range for the point estimate of TATE $\widehat{\text{TATE}} = \widehat{\text{zATE}} - \hat{\beta}_{vt}(\hat{E}_{S=1,W^{|Z}}[V] - E_{S=0}[V])$

V case toy example

OBSERVED DATA:	Trial sample: Treatment (n=200)	Trial sample: Control (n=200)	Trial sample: Full sample	Target population sample (n=10,000)
Covariates				
X = Years of education: mean (SD)	12.06 (1.64)	12.11 (1.58)	12.08 (1.61)	11.02 (1.52)
Z = Gender: percent female	49.50	50.50	50.00	19.86
V = Years smoked: mean (SD)	7.36 (2.57)	7.50 (2.45)	7.43 (2.51)	not observed
Outcome				
Y = Cigarettes per week: mean (SD)	97.42 (6.00)	101.80 (5.29)	99.61 (6.06)	not observed

Models fit to data:

$$\text{SATE model: } \hat{Y} = 120.31 - 2.02X - 4.36Z + 1.09V - 4.39T$$

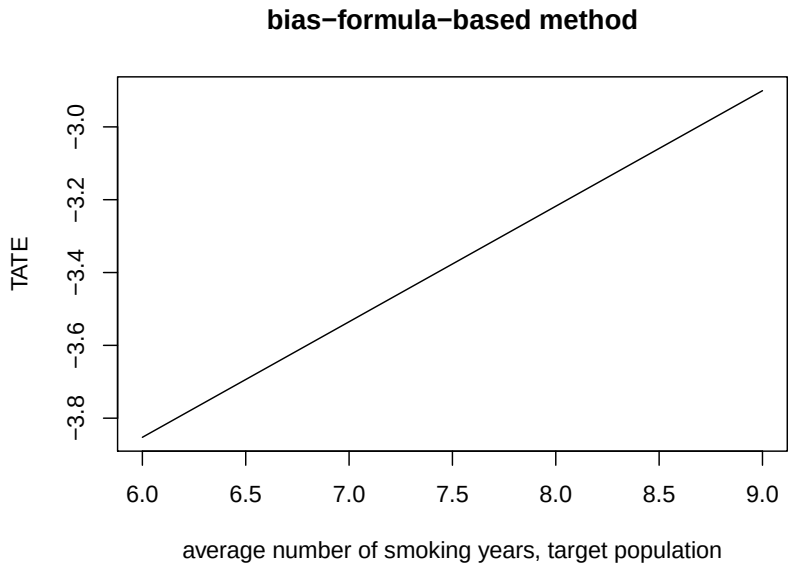
$$\text{effect mod. model: } \hat{Y} = 120.81 - 2.03X - 2.74Z + 0.93V - 5.11T - 3.27ZT + 0.32VT$$

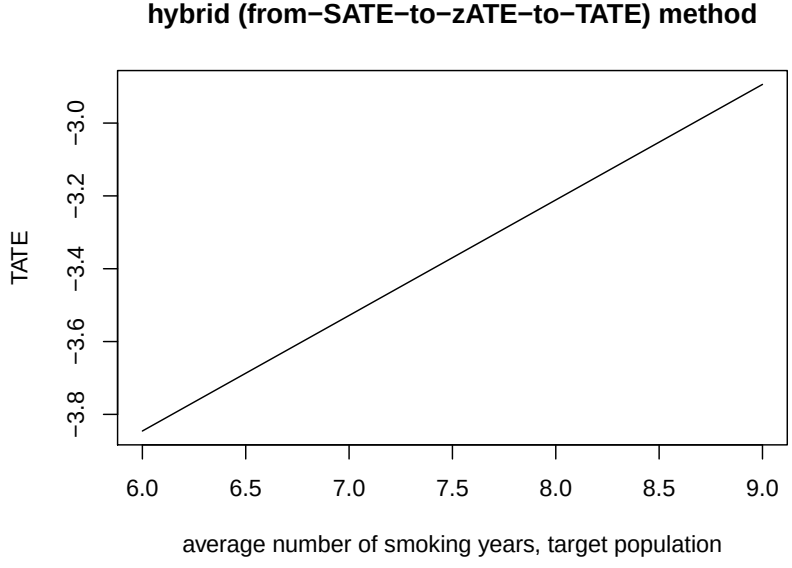
$\widehat{\text{SATE}} = -4.39$, 95% CI$=(-5.05, -3.73)$

V case toy example

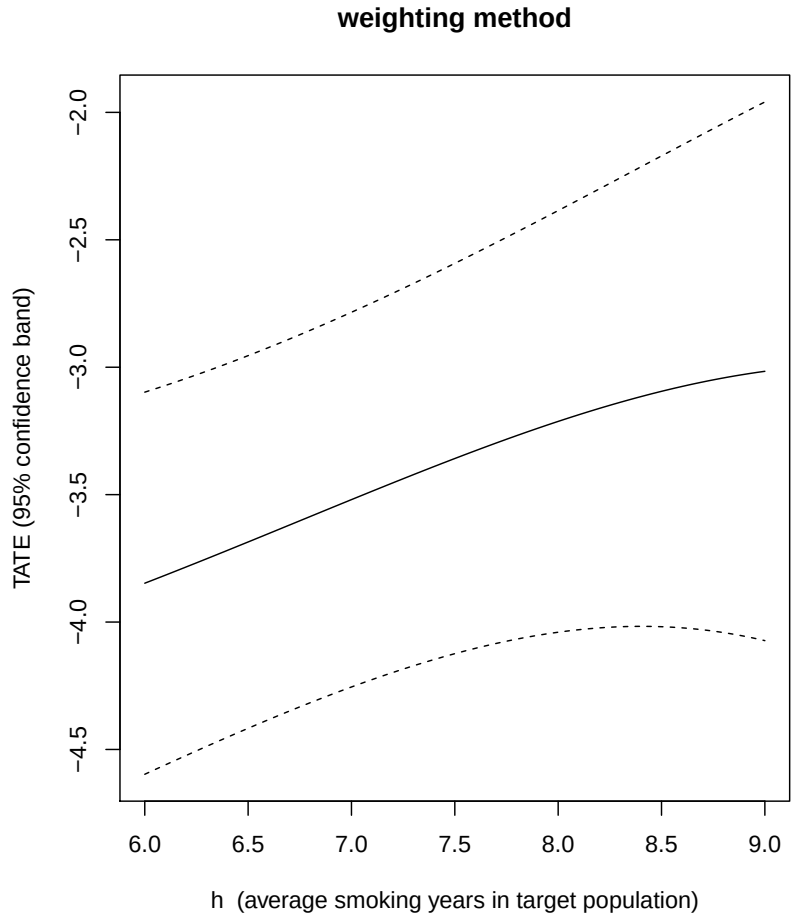
- Bias-formula-based sensitivity analysis
  - $\mathrm{E}_{S=0}[V]$ range specified to be 6-9 (smoking years)
- Weighting-based sensitivity analyses
  - $\frac{\mathrm{P}(S=0|Z)}{\mathrm{P}(S=1|Z)}$: two values for female and male participants
  - $\mathrm{P}(V|Z,S=1)$: Informed by trial data, assume and estimate a normal distribution conditional on gender
  - $\mathrm{P}(V|Z,S=0)$: In target population, suppose no reason to believe that women or men have smoked longer $\rightarrow$ specify normal distribution not conditional on gender, assuming variance equal to marginal variance from trial, with a moving mean ($\mathrm{E}_{S=0}[V]$) as the sensitivity parameter, also on the 6-9 range
- Hybrid method sensitivity analyses
  - $\widehat{\mathrm{zATE}} = -3.48$, 95% CI$= (-4.21, -2.76)$
  - $\mathrm{E}_{S=1,W|Z}[V] = 7.14$, 95% CI$= (6.86, 7.43)$
  - $\mathrm{E}_{S=0}[V]$ range specified to be 6-9 (smoking years)

V case toy example





three sensitivity analyses



U case: concerned about unobserved effect modification

Assumptions:

- Sample ignorability for treatment effects: $(Y^1 - Y^0) \perp S|Z, U$
- Overlap: the target population ranges of Z, U are covered by their ranges in the trial
- An additive model for the potential outcomes of the form $\mathrm{E}[Y_i^t] = \beta_0 + \beta_t t + \beta_{zt} Z_i t + \beta_{ut} U_i t + f_{xzu}(X_i, Z_i, U_i)$
- Additional assumption: U is independent of Z
  - the absence of U does not bias β_{zt}

$\Rightarrow$ **Definition:**

$U_{(z)} \equiv$ *remaining composite effect modifier* after accounting for Z

- Interpretation
  - a composite of residuals of unobserved effect modifiers
  - alternative: a natural variable, but have to assume it is the only unobserved effect modifier and is independent of Z (likely untrue)

U case: concerned about unobserved effect modification

$$\text{SATE} - \text{TATE} = \beta_{zt}(\text{E}_{S=1}[Z] - \text{E}_{S=0}[Z]) + \beta_{ut}(\text{E}_{S=1}[U_{(z)}] - \text{E}_{S=0}[U_{(z)}])$$

Bias-formula-based sensitivity analysis:

- Estimate SATE, $\text{E}_{S=1}[Z]$, $\text{E}_{S=0}[Z]$
- Estimate β_{zt} using regression analysis
  $Y = \beta_0 + \beta_t T + \beta_{zt} ZT + f_{xzv}(X, Z) + \epsilon$
- Specify plausible ranges for two sensitivity parameters β_{ut} and $\text{E}_{S=1}[U_{(z)}] - \text{E}_{S=0}[U_{(z)}]$
- Get a surface for the point estimate of TATE

U case: concerned about unobserved effect modification

Hybrid method sensitivity analysis for $U_{(z)}$:

- Weight trial sample using $W_i^{|Z_i}$ and estimate zATE

$$\begin{aligned} \text{zATE} &= \beta_t + \beta_{zt}\text{E}_{S=1,W^{|Z}}[Z] + \beta_{ut}\text{E}_{S=1,W^{|Z}}[U_{(z)}] \\ &= \beta_t + \beta_{zt}\text{E}_{S=0}[Z] + \beta_{ut}\text{E}_{S=1}[U_{(z)}] \\ \text{zATE} - \text{TATE} &= \beta_{ut}(\text{E}_{S=1}[U_{(z)}] - \text{E}_{S=0}[U_{(z)}]) \end{aligned}$$

- Specify plausible ranges for two sensitivity parameters β_{ut} and $\text{E}_{S=1}[U_{(z)}] - \text{E}_{S=0}[U_{(z)}]$
- Get three surfaces for TATE point estimate and confidence limits $\widehat{\text{TATE}} = \widehat{\text{zATE}} - \beta_{ut}(\text{E}_{S=1}[U_{(z)}] - \text{E}_{S=0}[U_{(z)}])$

Alternative: **Hybrid method sensitivity analysis for $U_{(xz)}$:**

$$\widehat{\text{TATE}} = \widehat{\text{xzATE}} - \beta_{ut}(\text{E}_{S=1}[U_{(xz)}] - \text{E}_{S=0}[U_{(xz)}])$$

U case toy example

OBSERVED DATA:	RCT sample			Target population sample (n=10,000)
	Treatment (n=200)	Control (n=200)	Full sample	
Covariates				
X = Years of education: mean (SD)	12.06 (1.64)	12.11 (1.58)	12.08 (1.61)	11.02 (1.52)
Z = Female gender: percent	49.50	50.50	50.00	19.86
U ?				
Outcome				
Y = Cigarettes per week: mean (SD)	97.42 (6.00)	101.80 (5.29)	99.61 (6.06)	

Models fit to data:

$$\text{SATE model: } \hat{Y} = 128.09 - 2.03X - 3.35Z - 4.53T$$

$$\text{effect modification model: } \hat{Y} = 127.50 - 2.04X - 1.98Z - 3.16T - 2.74ZT$$

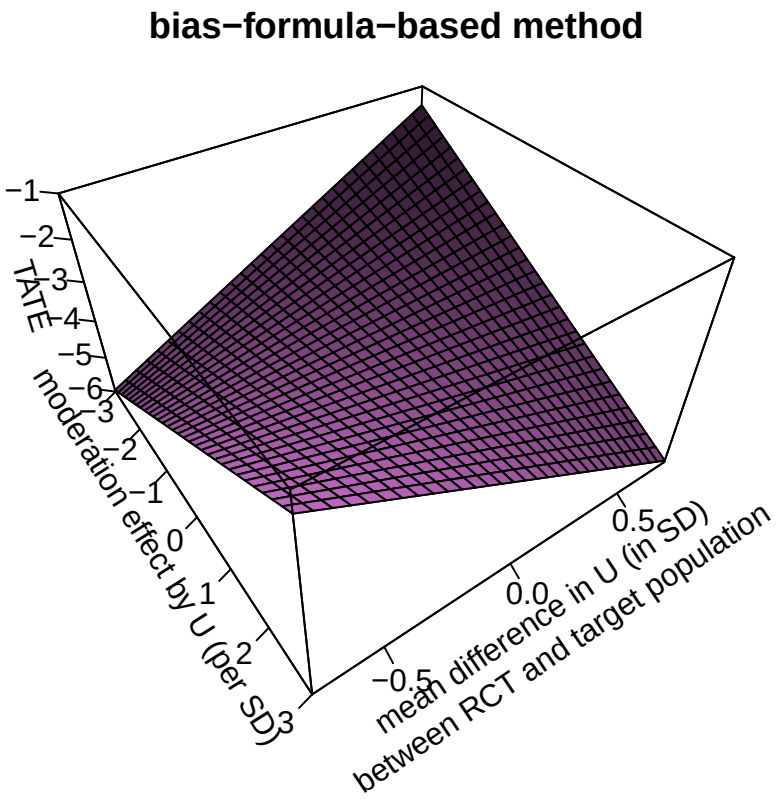
$\widehat{\text{SATE}} = -4.53$, 95% CI$=(-5.37, -3.69)$

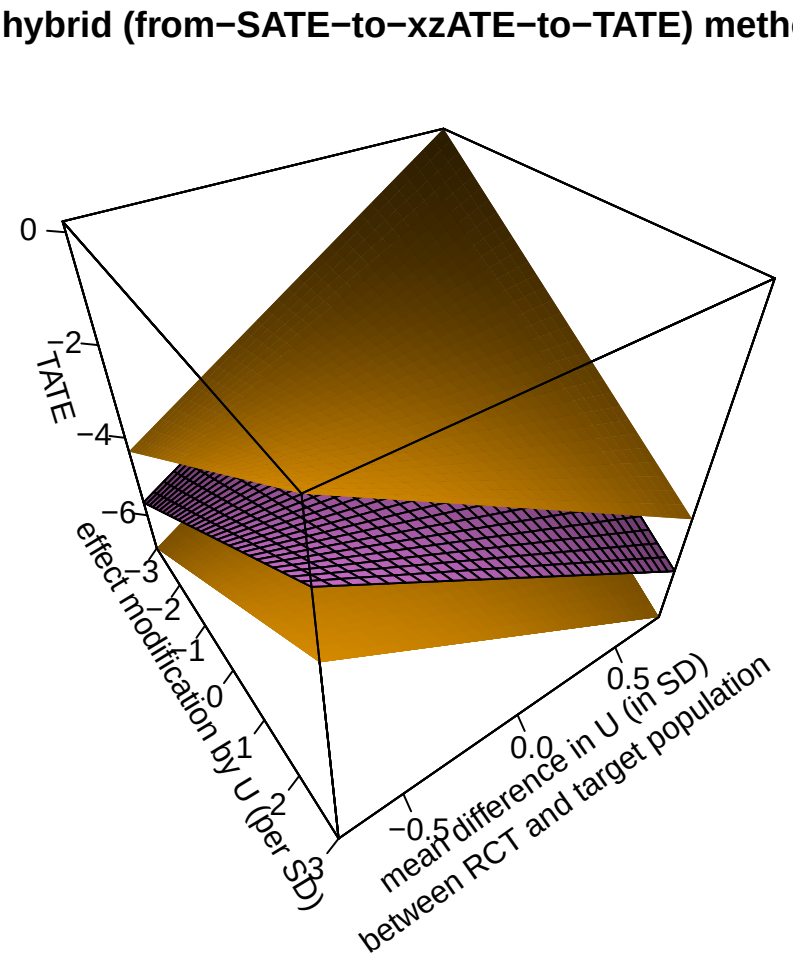
U case toy example

- Bias-formula-based sensitivity analysis for $U_{(z)}$
  - β_{ut} range: from -3 to 3 years per SD of $U_{(z)}$
  - $\mathrm{E}_{S=1}[U_{(z)}] - \mathrm{E}_{S=0}[U_{(z)}]$ range: from $-.7$ to .7 SD
- Hybrid method sensitivity analysis for $U_{(xz)}$
  - $\widehat{\text{xzATE}} = -3.39$, 95% CI$= (-4.72, -2.08)$
  - β_{ut} range: from -3 to 3 years per SD of $U_{(z)}$
  - $\mathrm{E}_{S=1}[U_{(xz)}] - \mathrm{E}_{S=0}[U_{(xz)}]$ range: from $-.7$ to .7 SD

U case toy example

two sensitivity analyses





Options for the two cases

	V case	U case
with some Z	bias-formula-based weighting-based hybrid (via zATE)	bias-formula-based [$U_{(z)}$] hybrid [$U_{(z)}$ or $U_{(xz)}$]
with no Z	bias-formula-based weighting-based	bias-formula-based [U]

Next steps

- a good data example
- capture uncertainty in the estimated parameter estimates for the bias-formula-based and hybrid methods for the V case
- for a binary outcome, investigate when the two methods based on the additive model fails
- extend to make use of target population outcome data when available
- extend V-case methods to address the situation when the scientist is concerned about a specific possible effect modifier that was not measured in the trial
- use a simulation-based approach that allows a more flexible outcome model

References

- Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. American Journal of Epidemiology, 172(1), 107-15. doi:10.1093/aje/kwq084
- Kern, H. L., Stuart, E. A., Hill, J. L., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. Journal of Research on Educational Effectiveness, 9(1), 103-127. doi:10.1080/19345747.2015.1060282