# A Simulation Study to Compare Causal Inference Methods for Point Exposures with Missing Confounders

**Luke Benz**[1]
Sebastien Haneuse[1] and Alexander Levis[2]

[1]Harvard T.H. Chan School of Public Health Department of Biostatistics

[2]Carnegie Mellon Department of Statistics & Data Science

**Joint Statistical Meetings 2023**
August 8, 2023

# The Problem

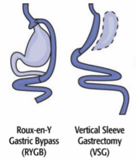
When using electronic health record (EHR) data to answer questions in comparative effectiveness:

- Treatment mechanism isn't random (**confounding**)
- Useful information may be absent (**missing data**)
- Surprisingly few papers attempt to formally address both confounding and missing data **simultaneously**

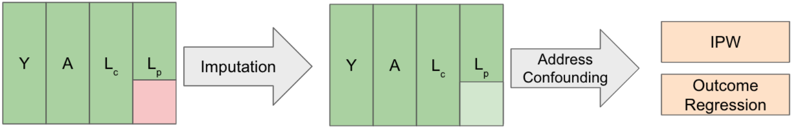
# A Motivating Example

Consider a study comparing two bariatric surgery procedures on 5 year weight loss outcomes

- **Treatment** ($A$): One of two bariatric surgery procedures
  - Roux-en-Y gastric bypass (RYGB) [Current "gold standard"]
  - Vertical sleeve gastrectomy (VSG) [Newer, less drastic procedure]
- **Outcome** ($Y$): % weight change 5 years post surgery
- **Confounders** ($L$):
  - **Fully Measured** ($L_c$): Baseline BMI, Race, Gender
  - **Partially Missing** ($L_p$): Comorbidities, Smoking Status



# Reasonable Approaches




- Several reasonable approaches could be conceived based on the following analysis pipeline
  1. (Multiple) Imputation to address missing data
  2. Adjustment for confounding on imputed dataset(s)
     - IPW
     - Outcome regression
  - Unclear when this strategy works well and when it doesn't
  - How do modeling choices affect this strategy?
  - Not always clear about what assumptions are being invoked
- Want a method that is
  - Clear in the assumptions being invoked
  - Flexible to model misspecification (e.g. doubly-robust)

# Notation

- Treatment: $A \in \mathcal{A}$
  - Point exposure
  - Finite number of treatments (e.g. $\mathcal{A}$ finite set)
- Outcome: $Y$
- Confounders: $L = (L_c, L_p)$
  - $L_c$: Observed for all subjects
  - $L_p$: Missing for some subjects
- Complete case indicator: $S$

- Counterfactual outcomes: $Y(a)$ for $a \in \mathcal{A}$
- Causal estimand of interest: $\mathbb{E}[Y(a)]$
  - Mean counterfactual outcome

# Assumptions

- Standard causal assumptions
  1. Consistency: $Y(A) = Y$
  2. No Unmeasured Confounding: $Y(a) \perp\!\!\!\perp A \mid L$, for all $a \in \mathcal{A}$
  3. Positivity: $P[A = a \mid L] \in (0, 1)$ for all $a \in \mathcal{A}$
- Under 1-3, $\mathbb{E}[Y(a)] = \mathbb{E}\Big[\mathbb{E}[Y|A, L]\Big]$
  - ...but we don't get to fully observe $L$!

# Assumptions

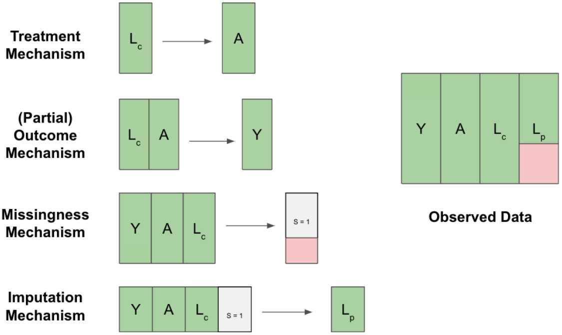
- Standard causal assumptions
  1. Consistency: $Y(A) = Y$
  2. No Unmeasured Confounding: $Y(a) \perp\!\!\!\perp A \mid L$, for all $a \in \mathcal{A}$
  3. Positivity: $P[A = a \mid L] \in (0, 1)$ for all $a \in \mathcal{A}$
- Under 1-3, $\mathbb{E}[Y(a)] = \mathbb{E}\Big[\mathbb{E}[Y|A, L]\Big]$
  - ...but we don't get to fully observe $L$!
- (Levis 2022) make the following missing data assumptions:
  4. Complete-case missing at random: $S \perp\!\!\!\perp L_p \mid L_c, A, Y$
  5. Complete-case positivity: $P[S = 1 \mid L_c, A, Y] \in (0, 1]$

# Levis Estimators

- (Levis 2022) derive 2 estimators when some confounders are partially missing, including one based on the efficient influence function (IF)
- IF estimator serves as benchmark w/ various theoretical guarantees
  - Doubly robust
  - Optimal asymptotic variance (in a non-parametric sense)
- Can be complex to compute; involves numerical integration techniques (e.g. Gaussian-quadrature)
- Based on novel factorization for the observed data likelihood

# Visual Intuition for Levis Factorization

**Factorization of observed data likelihood (pictures):**



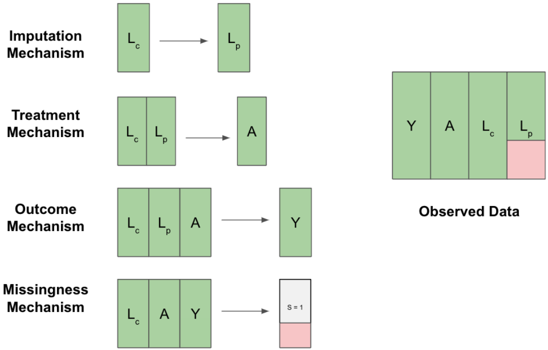

**Factorization of observed data likelihood (math):**

$$p(L_c)\underbrace{p(A \mid L_c)}_{\substack{\text{Treatment}\\ \text{Mechanism}}}\underbrace{p(Y \mid L_c, A)}_{\substack{\text{(Partial) Outcome}\\ \text{Mechanism}}}\underbrace{p(S \mid L_c, A, Y)}_{\substack{\text{Missingness}\\ \text{Mechanism}}}\underbrace{p(L_p \mid L_c, A, Y, S=1)^S}_{\substack{\text{Imputation}\\ \text{Mechanism}}}$$

**No component models depend on data we can't observe!**

# Visual Intuition for an Alternative Factorization

**Factorization of observed data likelihood (pictures):**




**Factorization of observed data likelihood (math):**

$$p(L_c)\underbrace{p(L_p \mid L_c)^s}_{\text{Imputation Mechanism}}\underbrace{p(A \mid L_c, L_p)}_{\text{Treatment Mechanism}}\underbrace{p(Y \mid L_c, L_p, A)}_{\text{Outcome Mechanism}}\underbrace{p(S \mid L_c, A, Y)}_{\text{Missingness Mechanism}}$$

**Component models depend on data we can't observe!**

# Simulation Study Outline

Conduct simulation study with following goals

1. Learn where ad-hoc approaches that consist of imputation plus some method that accounts for confounding are reasonable, and where they may breakdown

# Simulation Study Outline

Conduct simulation study with following goals

1. Learn where ad-hoc approaches that consist of imputation plus some method that accounts for confounding are reasonable, and where they may breakdown
2. Learn how the estimators proposed in (Levis 2022) perform when data are generated by an alternative factorization
   - True nuisance models are unknown

# Simulation Study Outline

Based on bariatric surgery motivating example

- 5,693 patients who underwent either of the two bariatric procedures of interest.
- Surgery at Kaiser Permanente Washington between 2008-2010.
- Complete information on gender, baseline BMI, and ethnicity.
- Comorbidity scores were only available for 4,344 patients.

# Simulation Study Outline

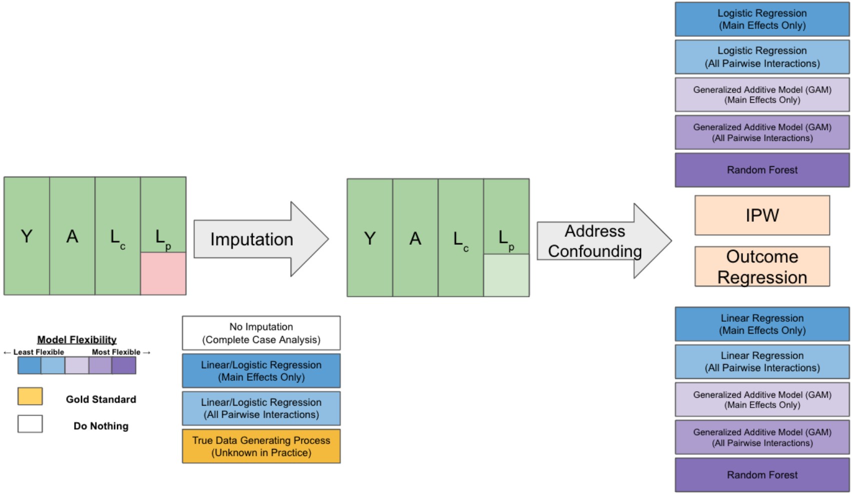
Based on bariatric surgery motivating example

- 5,693 patients who underwent either of the two bariatric procedures of interest.
- Surgery at Kaiser Permanente Washington between 2008-2010.
- Complete information on gender, baseline BMI, and ethnicity.
- Comorbidity scores were only available for 4,344 patients.

Kaiser data used to estimate “true” models for sampling

- Amplify certain relationships in different scenarios to get more interesting and complex relationships across confounders, treatment and outcome

# Modeling Choices for the Reasonable Analyst



# Modeling Choices: Levis Estimators

- Model types for nuisance functions:
  - **Treatment Model**: Logistic Regression
  - **(Partial) Outcome Model**: Linear Regression
  - **Missingness Model**: Linear Regression
  - **Imputation Model(s)**:
    - Comorbidities: Gamma GLM
    - Smoking: Logistic Regression
- Under **Levis Factorization** use "true" parametric models for nuisance functions (to establish baseline)
- Under **Alternative Factorization** where "true" parametric models for nuisance functions unknown, use:
  1. Same parametric models for nuisance functions as used under Levis factorization
  2. Flexible versions of these models via GAMs

# Summary

1. **Imputation**
   - Complete case analysis is severely biased.
   - Imputation method seems not to matter too much when Normal distribution is decent approximation for $\boldsymbol{L_p}$.
2. **Bias & Efficiency**
   - Sufficient model flexibility can overcome confounding bias due to model misspecification
   - Flexibility doesn't always come at the expense of efficiency
   - Model flexibility isn't a guarantee of unbiasedness
3. **Standard Methods**
   - Can perform well even with multiple missing confounders and amplified relationships between treatment/outcome/confounders
4. **Levis Estimators**
   - Levis IF estimator can be biased when nuisance functions misspecification.
   - Levis IF estimator with flexible modeling of nuisance functions can overcome bias due to misspecification.

# Takeaway

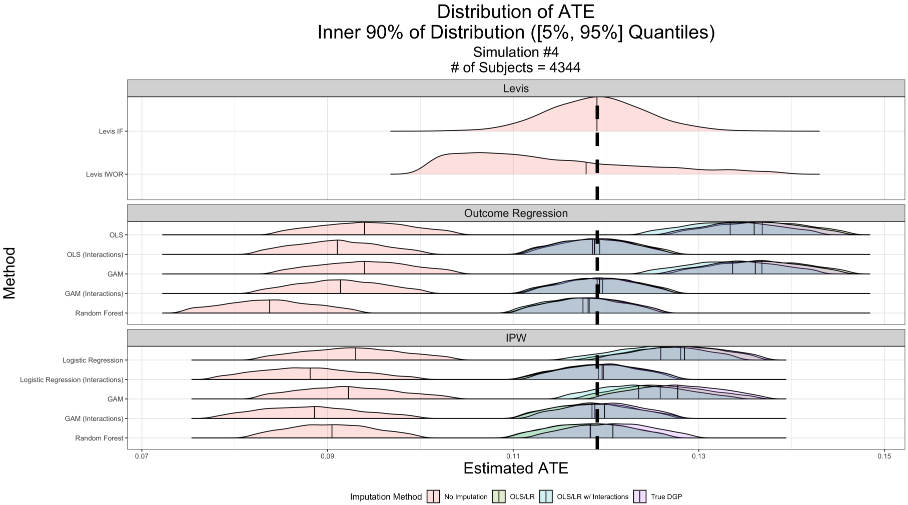
- Reasonable choices do reasonable things most of the time!
- In the absence of knowledge about missing data mechanisms, the work of (Levis 2022) may serve as a default for causal inference when handling confounding and missing data together.

# References

Levis, Alexander (2022). "Robust Methods for Causal Inference and Missing Data in Electronic Health Record-Based Comparative Effectiveness Research". PhD thesis. Boston, MA: Harvard University Graduate School of Arts and Sciences.

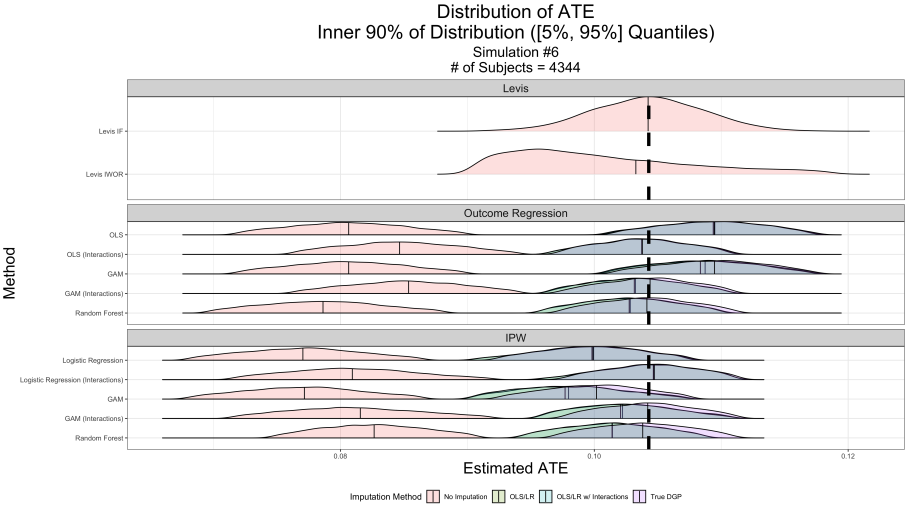
# Appendix

# Results



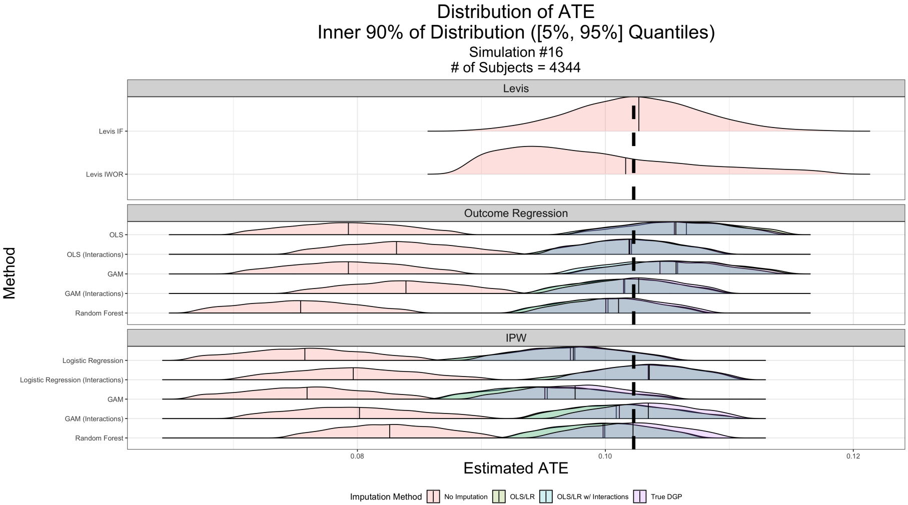

Levis Factorization, 1 Missing Confounder

# Results



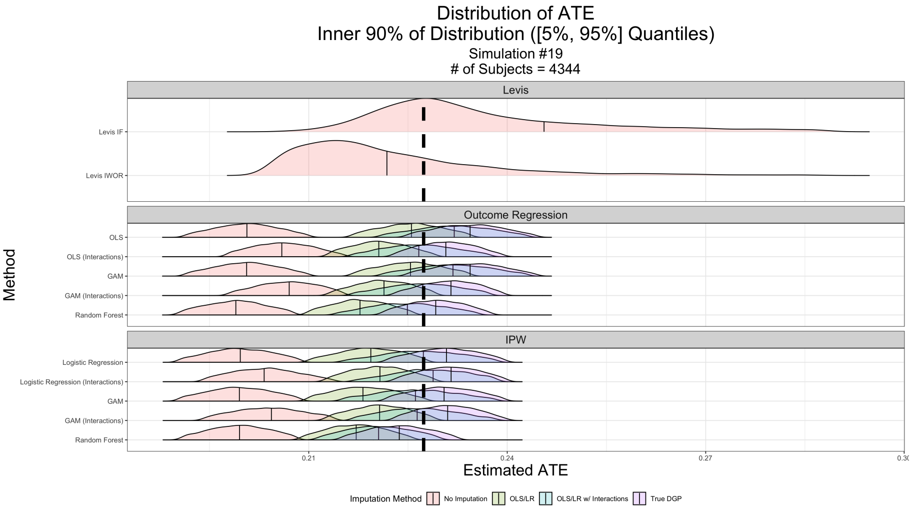

Levis Factorization, 1 Missing Confounder (More Skew)

# Results



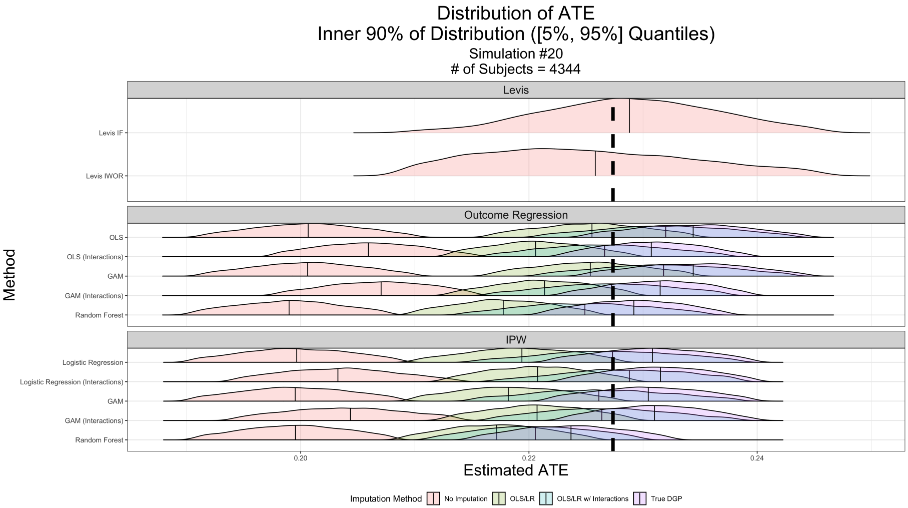

Levis Factorization, 2 Missing Confounders

# Results




Alternative Factorization
Levis Estimators: Parametric Models

# Results




Alternative Factorization

Levis Estimators: Semiparametric Models (GAMs)

# Formulation of Levis Influence Function Based Estimator

Table: Summary of nuisance functions for (Levis 2022) influence function based estimator. $P_{o'}$ denotes the joint distribution of the coarsened observed data, which consisents of $n$ replicates of $O' = (L_c, A, Y, S, SL_p)$.

| Nuisance Function | Definition | Description |
|---|---|---|
| $\eta(L_c, a)$ | $P_{o'}[A = a \mid L_c]$ | Treatment Mechanism |
| $\mu(y \mid L_c, A)$ | $P_{o'}[Y \leq y \mid L_c, A]$ | Outcome Distribution |
| $\pi(L_c, A, Y)$ | $P_{o'}[S = 1 \mid L_c, A, Y]$ | Missingness Mechanism |
| $\lambda(\ell_p \mid L_c, A, Y, S = 1)$ | $p_{o'}[\ell_p \mid L_c, A, Y, S = 1]$ | Imputation Model |

**Theorem 1 (Levis 2022)**: Under assumptions 1-5, the mean counterfactual $\mathbb{E}[Y(a)]$ is identified by the functional

$$\chi_a(P_{o'}) = \mathbb{E}_{P_{o'}}\left[\frac{S}{\pi(L_c, A, Y)}\xi(L_c, a; L_p)\right]$$

where

$$\xi(L_c, a; L_p) = \frac{\beta(L_c, a; L_p)}{\gamma(L_c, a; L_p)} = \frac{\int_{\mathcal{Y}} y\lambda(L_p \mid L_c, a, y, S)d\mu(y \mid L_c, a)}{\int_{\mathcal{Y}} \lambda(L_p \mid L_c, a, y, S)d\mu(y \mid L_c, a)}$$

# Formulation of Levis Influence Function Based Estimator

**Theorem 2 (Levis 2022)**: Under a non-parametric model for $P_{o'}$, the influence function of the mean counterfactual functional $\chi_a(P_{o'})$ is given by

$$\begin{aligned}
\dot{\chi}_a(O'; P_{o'}) &= \mathbb{E}_{P_{o'}}[\xi(L_c, a; L_p) \mid L_c, A = a, Y, S = 1] - \chi_a(P_{o'}) \\
&+ \frac{S}{\pi(L_c, A, Y)} \left\{ \xi(L_c, a; L_p) - \mathbb{E}_{P_{o'}}[\xi(L_c, a; L_p) \mid L_c, A = a, Y, S = 1] \right\} \\
&+ \frac{\mathbb{1}\{A = a\}}{\eta(L_c, a)} \mathbb{E}_{P_{o'}}[\epsilon_a(L_c, Y; L_p) \mid L_c, A = a, Y, S = 1] \\
&+ \frac{S}{\pi(L_c, A, Y)} \frac{\mathbb{1}\{A = a\}}{\eta(L_c, a)} \left\{ \epsilon_a(L_c, Y; L_p) - \mathbb{E}_{P_{o'}}[\epsilon_a(L_c, Y; L_p) \mid L_c, A = a, Y, S = 1] \right\}
\end{aligned}$$

where

$$\tau(L_c; L_p) = \sum_{a'=0}^{1} \eta(L_c, a') \gamma(L_c, a'; L_p)$$

$$\epsilon_a(L_c Y; L_p) = \frac{\tau(L_c; L_p)}{\gamma(L_c, a; L_p)} \{Y - \chi(L_c, a; L_p)\}$$

$$\mathbb{E}_{P_{o'}}[h(L_c, A, Y; L_p) \mid L_c, A, Y, S = 1] = \int_{\mathcal{L}_p} h(L_c, A, Y; \ell_p) \lambda(\ell_p \mid L_c, A, Y, S) d\nu(\ell_p)$$

($\nu$ is dominating measure for density $\lambda$)

# Formulation of Levis Influence Function Based Estimator

Using these theorems, (Levis 2022) propose the following one-step influence function-based estimator of $\mathbb{E}[Y(a)]$

$$\hat{\chi}_a = \chi_a(\hat{P}_{o'}) + \frac{1}{n}\sum_{i=1}^{n} \dot{\chi}_a(O_i'; \hat{P}_{o'})$$

Note that this estimator requires plug-in estimates for all four nuisance functions summarized in Table 1.