Causal Inference Methods for Infectious Disease Prevention Trials

Brian Richardson

Department of Biostatistics, University of North Carolina at Chapel Hill

November 03, 2025

Scan for slides

Outline

- **Project 1**: Addressing Confounding and Continuous Exposure Measurement Error Using Corrected Score Functions

Outline

- **Project 1**: Addressing Confounding and Continuous Exposure Measurement Error Using Corrected Score Functions
- **Project 2**: Social Distancing to Reduce Transmission of Influenza-Like-Illness on College Campuses: the eX-FLU Trial

Outline

- **Project 1**: Addressing Confounding and Continuous Exposure Measurement Error Using Corrected Score Functions
- **Project 2**: Social Distancing to Reduce Transmission of Influenza-Like-Illness on College Campuses: the eX-FLU Trial
- **Project 3**: Causal Inference from Cluster Randomized Trials with Differential Nonresponse

Project 1 : Addressing Confounding and Continuous Exposure Measurement Error Using Corrected Score Functions

Brian Richardson, Brian Blette, Peter Gilbert, Michael Hudgens (2025)

Motivation: HVTN 505 trial

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 NOVEMBER 28, 2013 VOL. 369 NO. 22

Efficacy Trial of a DNA/rAd5 HIV-1 Preventive Vaccine

Scott M. Hammer, M.D., Magdalena E. Sobieszczyk, M.D., M.P.H., Holly Janes, Ph.D., Shelly T. Karuna, M.D., Mark J. Mulligan, M.D., Doug Grove, M.S., Beryl A. Koblin, Ph.D., Susan P. Buchbinder, M.D., Michael C. Keefer, M.D., Georgia D. Tomaras, Ph.D., Nicole Frahm, Ph.D., John Hural, Ph.D., Chuka Anude, M.D., Ph.D., Barney S. Graham, M.D., Ph.D., Mary E. Enama, M.A., P.A.-C., Elizabeth Adams, M.D., Edwin DeJesus, M.D., Richard M. Novak, M.D., Ian Frank, M.D., Carter Bentley, Ph.D., Shelly Ramirez, M.A., Rong Fu, M.S., Richard A. Koup, M.D., John R. Mascola, M.D., Gary J. Nabel, M.D., Ph.D., David C. Montefiori, Ph.D., James Kublin, M.D., M.P.H., M. Juliana McElrath, M.D., Ph.D., Lawrence Corey, M.D., and Peter B. Gilbert, Ph.D., for the HVTN 505 Study Team*

- **HVTN 505 trial**: trial of a preventive HIV vaccine

Motivation: HVTN 505 trial

The NEW ENGLAND JOURNAL of MEDICINE

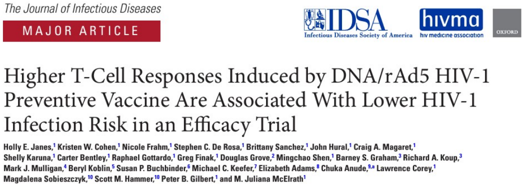
ESTABLISHED IN 1812 NOVEMBER 28, 2013 VOL. 369 NO. 22

Efficacy Trial of a DNA/rAd5 HIV-1 Preventive Vaccine

Scott M. Hammer, M.D., Magdalena E. Sobieszczyk, M.D., M.P.H., Holly Janes, Ph.D., Shelly T. Karuna, M.D., Mark J. Mulligan, M.D., Doug Grove, M.S., Beryl A. Koblin, Ph.D., Susan P. Buchbinder, M.D., Michael C. Keefer, M.D., Georgia D. Tomaras, Ph.D., Nicole Frahm, Ph.D., John Hural, Ph.D., Chuka Anude, M.D., Ph.D., Barney S. Graham, M.D., Ph.D., Mary E. Enama, M.A., P.A.-C., Elizabeth Adams, M.D., Edwin DeJesus, M.D., Richard M. Novak, M.D., Ian Frank, M.D., Carter Bentley, Ph.D., Shelly Ramirez, M.A., Rong Fu, M.S., Richard A. Koup, M.D., John R. Mascola, M.D., Gary J. Nabel, M.D., Ph.D., David C. Montefiori, Ph.D., James Kublin, M.D., M.P.H., M. Juliana McElrath, M.D., Ph.D., Lawrence Corey, M.D., and Peter B. Gilbert, Ph.D., for the HVTN 505 Study Team*

- **HVTN 505 trial**: trial of a preventive HIV vaccine
- Stopped early after reaching predetermined cutoffs for efficacy futility (Hammer et al., 2013)

Motivation: HVTN 505 trial


The Journal of Infectious Diseases
MAJOR ARTICLE

Higher T-Cell Responses Induced by DNA/rAd5 HIV-1 Preventive Vaccine Are Associated With Lower HIV-1 Infection Risk in an Efficacy Trial

Holly E. Janes,[1] Kristen W. Cohen,[1] Nicole Frahm,[1] Stephen C. De Rosa,[1] Brittany Sanchez,[1] John Hural,[1] Craig A. Magaret,[1] Shelly Karuna,[1] Carter Bentley,[1] Raphael Gottardo,[1] Greg Finak,[1] Douglas Grove,[2] Mingchao Shen,[1] Barney S. Graham,[3] Richard A. Koup,[3] Mark J. Mulligan,[4] Beryl Koblin,[5] Susan P. Buchbinder,[6] Michael C. Keefer,[7] Elizabeth Adams,[8] Chuka Anude,[9,a] Lawrence Corey,[1] Magdalena Sobieszczyk,[10] Scott M. Hammer,[10] Peter B. Gilbert,[1] and M. Juliana McElrath[1]

- Several biomarkers correlated with HIV among vaccine recipients (Janes et al., 2017; Fong et al., 2018; Neidich et al., 2019)

Motivation: HVTN 505 trial

The Journal of Infectious Diseases

MAJOR ARTICLE

Higher T-Cell Responses Induced by DNA/rAd5 HIV-1 Preventive Vaccine Are Associated With Lower HIV-1 Infection Risk in an Efficacy Trial

Holly E. Janes,[1] Kristen W. Cohen,[1] Nicole Frahm,[1] Stephen C. De Rosa,[1] Brittany Sanchez,[1] John Hural,[1] Craig A. Magaret,[1] Shelly Karuna,[1] Carter Bentley,[1] Raphael Gottardo,[1] Greg Finak,[1] Douglas Grove,[2] Mingchao Shen,[1] Barney S. Graham,[3] Richard A. Koup,[3] Mark J. Mulligan,[4] Beryl Koblin,[5] Susan P. Buchbinder,[6] Michael C. Keefer,[7] Elizabeth Adams,[8] Chuka Anude,[9,a] Lawrence Corey,[1] Magdalena Sobieszczyk,[10] Scott M. Hammer,[10] Peter B. Gilbert,[1] and M. Juliana McElrath[1]

- Several biomarkers correlated with HIV among vaccine recipients (Janes et al., 2017; Fong et al., 2018; Neidich et al., 2019)
- Is there a **causal** relationship between these biomarkers and HIV?

Motivation: HVTN 505 trial

The Journal of Infectious Diseases

MAJOR ARTICLE

IDSA Infectious Diseases Society of America hivma hiv medicine association OXFORD

Higher T-Cell Responses Induced by DNA/rAd5 HIV-1 Preventive Vaccine Are Associated With Lower HIV-1 Infection Risk in an Efficacy Trial

Holly E. Janes,[1] Kristen W. Cohen,[1] Nicole Frahm,[1] Stephen C. De Rosa,[1] Brittany Sanchez,[1] John Hural,[1] Craig A. Magaret,[1] Shelly Karuna,[1] Carter Bentley,[1] Raphael Gottardo,[1] Greg Finak,[1] Douglas Grove,[2] Mingchao Shen,[1] Barney S. Graham,[3] Richard A. Koup,[3] Mark J. Mulligan,[4] Beryl Koblin,[5] Susan P. Buchbinder,[6] Michael C. Keefer,[7] Elizabeth Adams,[8] Chuka Anude,[9,a] Lawrence Corey,[1] Magdalena Sobieszczyk,[10] Scott M. Hammer,[10] Peter B. Gilbert,[1] and M. Juliana McElrath[1]

- Several biomarkers correlated with HIV among vaccine recipients (Janes et al., 2017; Fong et al., 2018; Neidich et al., 2019)
- Is there a **causal** relationship between these biomarkers and HIV?
- Biomarker-HIV relationship is **confounded**

Motivation: HVTN 505 trial

The Journal of Infectious Diseases

MAJOR ARTICLE

IDSA Infectious Diseases Society of America · hivma hiv medicine association · OXFORD

Higher T-Cell Responses Induced by DNA/rAd5 HIV-1 Preventive Vaccine Are Associated With Lower HIV-1 Infection Risk in an Efficacy Trial

Holly E. Janes,[1] Kristen W. Cohen,[1] Nicole Frahm,[1] Stephen C. De Rosa,[1] Brittany Sanchez,[1] John Hural,[1] Craig A. Magaret,[1] Shelly Karuna,[1] Carter Bentley,[1] Raphael Gottardo,[1] Greg Finak,[1] Douglas Grove,[2] Mingchao Shen,[1] Barney S. Graham,[3] Richard A. Koup,[3] Mark J. Mulligan,[4] Beryl Koblin,[5] Susan P. Buchbinder,[6] Michael C. Keefer,[7] Elizabeth Adams,[8] Chuka Anude,[9,a] Lawrence Corey,[1] Magdalena Sobieszczyk,[10] Scott M. Hammer,[10] Peter B. Gilbert,[1] and M. Juliana McElrath[1]

- Several biomarkers correlated with HIV among vaccine recipients (Janes et al., 2017; Fong et al., 2018; Neidich et al., 2019)
- Is there a **causal** relationship between these biomarkers and HIV?
- Biomarker-HIV relationship is **confounded**
- Biomarkers are **measured with error**

Goal

To estimate the **causal effect** of a continuous exposure on an outcome when

Goal

To estimate the **causal effect** of a continuous exposure on an outcome when

(i) The exposure-outcome relationship is potentially **confounded**

Goal

To estimate the **causal effect** of a continuous exposure on an outcome when

(i) The exposure-outcome relationship is potentially **confounded**

(ii) The exposure is **measured with error**

What Do We Mean by Causal Effect?

"What would be the risk of HIV if, possibly counter to fact, somebody were to have biomarker level $\mathbf{a}$?" (Neyman et al., 1935; Rubin, 1974; Holland, 1986; Hernan and Robins, 2024)

What Do We Mean by Causal Effect?

"What would be the risk of HIV if, possibly counter to fact, somebody were to have biomarker level $\mathbf{a}$?" (Neyman et al., 1935; Rubin, 1974; Holland, 1986; Hernan and Robins, 2024)

$$\underbrace{\mathbf{a}}_{\text{biomarker}}$$

What Do We Mean by Causal Effect?

"What would be the risk of HIV if, possibly counter to fact, somebody were to have biomarker level **a**?" (Neyman et al., 1935; Rubin, 1974; Holland, 1986; Hernan and Robins, 2024)

$$\underbrace{\mathbf{a}}_{\text{biomarker}} \quad \Rightarrow \quad \underbrace{Y(\mathbf{a})}_{\text{potential HIV status if } \mathbf{a}}$$

What Do We Mean by Causal Effect?

"What would be the risk of HIV if, possibly counter to fact, somebody were to have biomarker level **a**?" (Neyman et al., 1935; Rubin, 1974; Holland, 1986; Hernan and Robins, 2024)

$$\underbrace{\mathbf{a}}_{\text{biomarker}} \Rightarrow \underbrace{Y(\mathbf{a})}_{\text{potential HIV status if } \mathbf{a}} \Rightarrow \underbrace{\mathrm{E}\{Y(\mathbf{a})\}}_{\text{potential HIV risk if } \mathbf{a}}$$

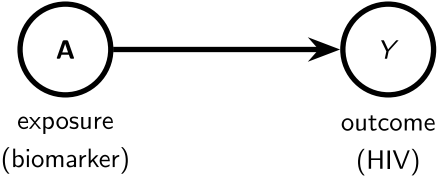
What Do We Mean by Causal Effect?

"What would be the risk of HIV if, possibly counter to fact, somebody were to have biomarker level **a**?" (Neyman et al., 1935; Rubin, 1974; Holland, 1986; Hernan and Robins, 2024)

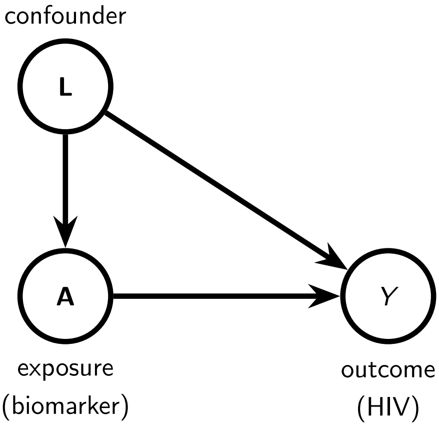
$$\underbrace{\mathbf{a}}_{\text{biomarker}} \Rightarrow \underbrace{Y(\mathbf{a})}_{\text{potential HIV status if } \mathbf{a}} \Rightarrow \underbrace{\mathrm{E}\{Y(\mathbf{a})\}}_{\text{potential HIV risk if } \mathbf{a}}$$

Estimand (**dose-response surface**): $\eta(\mathbf{a}) \equiv \mathrm{E}\{Y(\mathbf{a})\}$ for $\mathbf{a} \in \mathcal{A}$

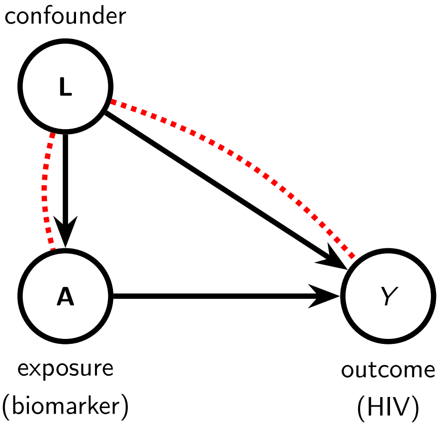
Confounding



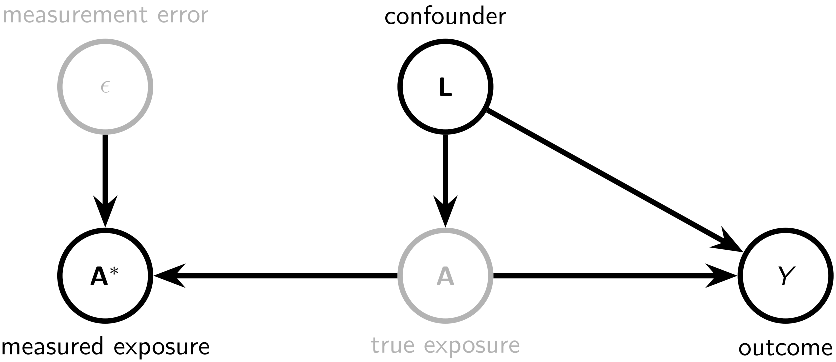
Confounding



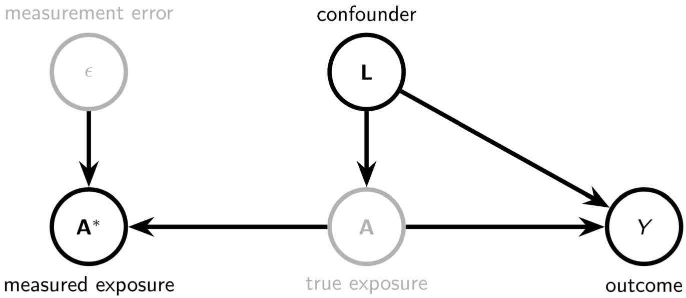
Confounding



Measurement Error

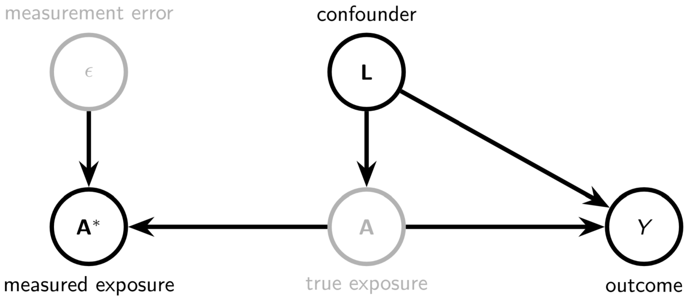


Notation



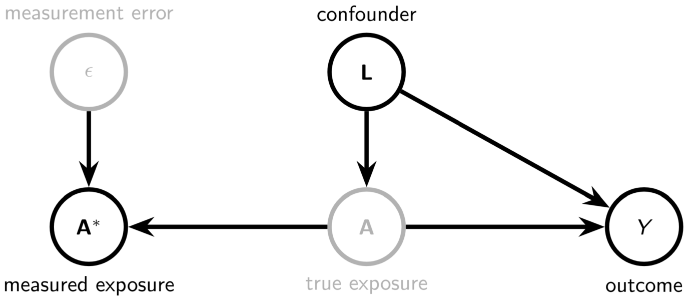
measurement error
confounder
ϵ
L
A^*
A
Y
measured exposure
true exposure
outcome

Notation



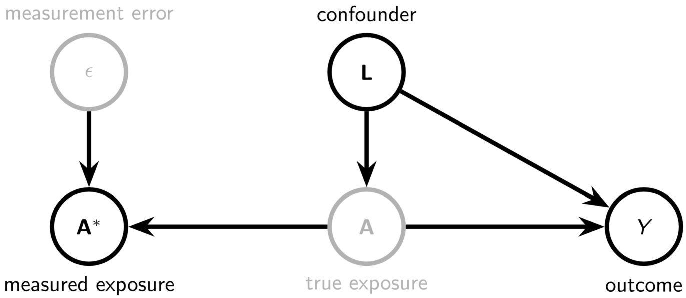
- True exposure: $\mathbf{A} = (A_1, \ldots, A_m)$

Notation



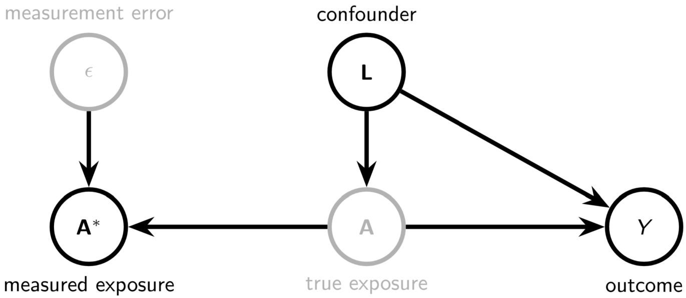
- True exposure: $\mathbf{A} = (A_1, \ldots, A_m)$
- Measurement error: $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$

Notation



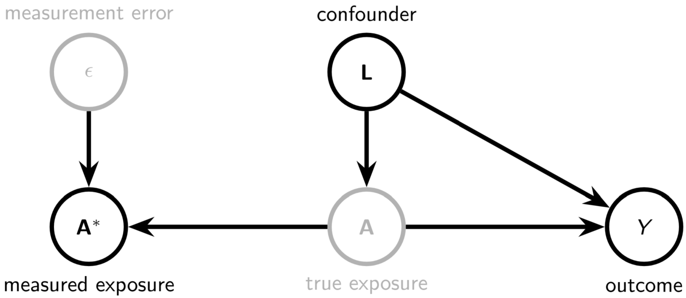
- True exposure: $\mathbf{A} = (A_1, \ldots, A_m)$
- Measurement error: $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$
- Measured exposure: $\mathbf{A}^* = (A_1^*, \ldots, A_m^*) = \mathbf{A} + \epsilon$

Notation



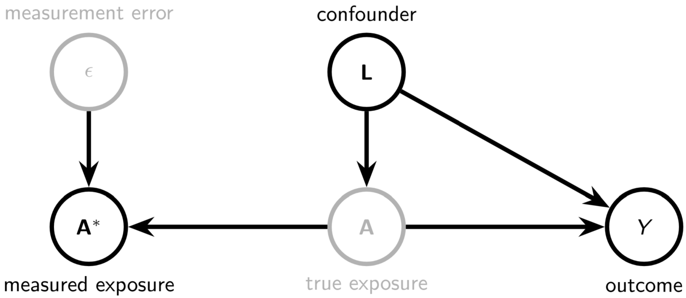
- Measured exposure: $\mathbf{A}^* = (A_1^*, \ldots, A_m^*) = \mathbf{A} + \epsilon$
- True exposure: $\mathbf{A} = (A_1, \ldots, A_m)$
- Measurement error: $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$
- Potential outcome: $Y(\mathbf{a})$

Notation



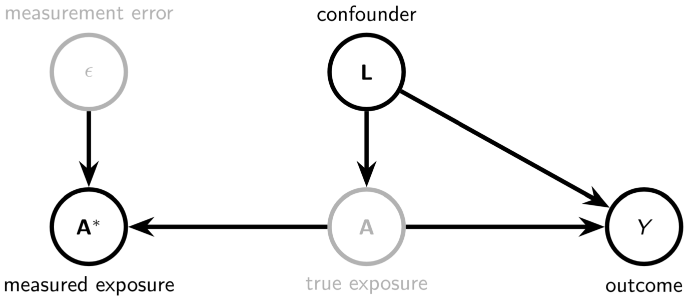
- True exposure: $\mathbf{A} = (A_1, \ldots, A_m)$
- Measurement error: $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$
- Potential outcome: $Y(\mathbf{a})$
- Measured exposure: $\mathbf{A}^* = (A_1^*, \ldots, A_m^*) = \mathbf{A} + \epsilon$
- Observed outcome: Y

Notation



- True exposure: $\mathbf{A} = (A_1, \ldots, A_m)$
- Measurement error: $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$
- Potential outcome: $Y(\mathbf{a})$
- Measured exposure: $\mathbf{A}^* = (A_1^*, \ldots, A_m^*) = \mathbf{A} + \epsilon$
- Observed outcome: Y
- Confounders: $\mathbf{L} = (L_1, L_2, ..., L_p)$

Notation



- True exposure: $\mathbf{A} = (A_1, \ldots, A_m)$
- Measurement error: $\epsilon = (\epsilon_1, \ldots, \epsilon_m)$
- Potential outcome: $Y(\mathbf{a})$

- Measured exposure: $\mathbf{A}^* = (A_1^*, \ldots, A_m^*) = \mathbf{A} + \epsilon$
- Observed outcome: Y
- Confounders: $\mathbf{L} = (L_1, L_2, ..., L_p)$
- **Observe**: iid copies of $(Y_i, \mathbf{L}_i, \mathbf{A}_i^*)$

Assumptions

Observe: iid copies of $(Y_i, \mathbf{L}_i, \mathbf{A}_i^*)$

Assumptions

Observe: iid copies of $(Y_i, \mathbf{L}_i, \mathbf{A}_i^*)$

Estimand: dose-response surface $\eta(\mathbf{a}) \equiv \mathrm{E}\{Y(\mathbf{a})\}$ for $\mathbf{a} \in \mathcal{A}$

Assumptions

Observe: iid copies of $(Y_i, \mathbf{L}_i, \mathbf{A}_i^*)$

Estimand: dose-response surface $\eta(\mathbf{a}) \equiv \mathsf{E}\{Y(\mathbf{a})\}$ for $\mathbf{a} \in \mathcal{A}$

(i) **Causal consistency**: $Y = Y(\mathbf{a})$ when $\mathbf{A} = \mathbf{a}$

Assumptions

Observe: iid copies of $(Y_i, \mathbf{L}_i, \mathbf{A}_i^*)$

Estimand: dose-response surface $\eta(\mathbf{a}) \equiv \mathrm{E}\{Y(\mathbf{a})\}$ for $\mathbf{a} \in \mathcal{A}$

(i) **Causal consistency**: $Y = Y(\mathbf{a})$ when $\mathbf{A} = \mathbf{a}$

(ii) **Conditional exchangeability**: $Y(\mathbf{a}) \perp\!\!\!\perp \mathbf{A}|\mathbf{L}$ for all $\mathbf{a} \in \mathcal{A}$

Assumptions

Observe: iid copies of $(Y_i, \mathbf{L}_i, \mathbf{A}_i^*)$

Estimand: dose-response surface $\eta(\mathbf{a}) \equiv \mathrm{E}\{Y(\mathbf{a})\}$ for $\mathbf{a} \in \mathcal{A}$

(i) **Causal consistency**: $Y = Y(\mathbf{a})$ when $\mathbf{A} = \mathbf{a}$

(ii) **Conditional exchangeability**: $Y(\mathbf{a}) \perp\!\!\!\perp \mathbf{A}|\mathbf{L}$ for all $\mathbf{a} \in \mathcal{A}$

(iii) **Positivity**: $f_{\mathbf{A}|\mathbf{L}}(\mathbf{a}|\mathbf{l}) > 0$ for all $\mathbf{l}$ such that $f_{\mathbf{L}}(\mathbf{l}) > 0$ and for all $\mathbf{a} \in \mathcal{A}$

Assumptions

Observe: iid copies of $(Y_i, \mathbf{L}_i, \mathbf{A}_i^*)$

Estimand: dose-response surface $\eta(\mathbf{a}) \equiv \mathrm{E}\{Y(\mathbf{a})\}$ for $\mathbf{a} \in \mathcal{A}$

(i) **Causal consistency**: $Y = Y(\mathbf{a})$ when $\mathbf{A} = \mathbf{a}$

(ii) **Conditional exchangeability**: $Y(\mathbf{a}) \perp\!\!\!\perp \mathbf{A}|\mathbf{L}$ for all $\mathbf{a} \in \mathcal{A}$

(iii) **Positivity**: $f_{\mathbf{A}|\mathbf{L}}(\mathbf{a}|\mathbf{l}) > 0$ for all $\mathbf{l}$ such that $f_{\mathbf{L}}(\mathbf{l}) > 0$ and for all $\mathbf{a} \in \mathcal{A}$

(iv) **Classical additive measurement error**: $\epsilon \sim \mathcal{N}_m(0, \mathbf{\Sigma}_{me})$ and $\epsilon \perp\!\!\!\perp (Y, \mathbf{L}, \mathbf{A})$

Assumptions

Observe: iid copies of $(Y_i, \mathbf{L}_i, \mathbf{A}_i^*)$

Estimand: dose-response surface $\eta(\mathbf{a}) \equiv \mathrm{E}\{Y(\mathbf{a})\}$ for $\mathbf{a} \in \mathcal{A}$

(i) **Causal consistency**: $Y = Y(\mathbf{a})$ when $\mathbf{A} = \mathbf{a}$

(ii) **Conditional exchangeability**: $Y(\mathbf{a}) \perp\!\!\!\perp \mathbf{A} | \mathbf{L}$ for all $\mathbf{a} \in \mathcal{A}$

(iii) **Positivity**: $f_{\mathbf{A}|\mathbf{L}}(\mathbf{a}|\mathbf{l}) > 0$ for all $\mathbf{l}$ such that $f_{\mathbf{L}}(\mathbf{l}) > 0$ and for all $\mathbf{a} \in \mathcal{A}$

(iv) **Classical additive measurement error**: $\epsilon \sim \mathcal{N}_m(0, \mathbf{\Sigma}_{me})$ and $\epsilon \perp\!\!\!\perp (Y, \mathbf{L}, \mathbf{A})$

(v) **Known or estimable measurement error variance $\mathbf{\Sigma}_{me}$**

Addressing Confounding Alone

How can we address confounding in the absence of measurement error?

Addressing Confounding Alone

How can we address confounding in the absence of measurement error?

- 3 classical methods:

Addressing Confounding Alone

How can we address confounding in the absence of measurement error?

- 3 classical methods:
    1. G-Formula (Robins, 1986; Hernan and Robins, 2024)

Addressing Confounding Alone

How can we address confounding in the absence of measurement error?

- 3 classical methods:
    1. G-Formula (Robins, 1986; Hernan and Robins, 2024)
    2. Inverse Probability Weighting (IPW) (Rosenbaum and Rubin, 1983; Hernan and Robins, 2024)

Addressing Confounding Alone

How can we address confounding in the absence of measurement error?

- 3 classical methods:
  1. G-Formula (Robins, 1986; Hernan and Robins, 2024)
  2. Inverse Probability Weighting (IPW) (Rosenbaum and Rubin, 1983; Hernan and Robins, 2024)
  3. Doubly Robust Method (DR) (Kang and Schafer, 2007; Zhang et al., 2016)

Addressing Confounding Alone

How can we address confounding in the absence of measurement error?

- 3 classical methods:
  1. G-Formula (Robins, 1986; Hernan and Robins, 2024)
  2. Inverse Probability Weighting (IPW) (Rosenbaum and Rubin, 1983; Hernan and Robins, 2024)
  3. Doubly Robust Method (DR) (Kang and Schafer, 2007; Zhang et al., 2016)
- These can all be framed as **M-estimators** (Stefanski and Boos, 2002)

Crash Course on M-Estimation

Score Function: a function of the observed data and the parameter of interest

$$\underbrace{\boldsymbol{\Psi}_0}_{\text{score fun.}} \left(\qquad ; \qquad \right)$$

Crash Course on M-Estimation

Score Function: a function of the observed data and the parameter of interest

$$\underbrace{\boldsymbol{\Psi}_0}_{\text{score fun.}} \left(\underbrace{Y, \mathbf{L}, \mathbf{A}}_{\text{data}} ; \qquad \right)$$

Crash Course on M-Estimation

Score Function: a function of the observed data and the parameter of interest

$$\underbrace{\boldsymbol{\Psi}_0}_{\text{score fun.}} \left(\underbrace{Y, \mathbf{L}, \mathbf{A}}_{\text{data}} ; \underbrace{\theta}_{\text{param.}} \right)$$

Crash Course on M-Estimation

Score Function: a function of the observed data and the parameter of interest

$$\underbrace{\boldsymbol{\Psi}_0}_{\text{score fun.}} \left(\underbrace{Y, \mathbf{L}, \mathbf{A}}_{\text{data}} ; \underbrace{\theta}_{\text{param.}} \right)$$

that is **unbiased**, meaning

Crash Course on M-Estimation

Score Function: a function of the observed data and the parameter of interest

$$\underbrace{\boldsymbol{\Psi}_0}_{\text{score fun.}}\ (\underbrace{Y, \mathbf{L}, \mathbf{A}}_{\text{data}}\ ;\ \underbrace{\theta}_{\text{param.}})$$

that is **unbiased**, meaning

$$\mathrm{E}\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}; \underbrace{\theta_0}_{\text{true}})\} = \mathbf{0}.$$

Crash Course on M-Estimation

Score Function: a function of the observed data and the parameter of interest

$$\underbrace{\boldsymbol{\Psi}_0}_{\text{score fun.}} \; (\underbrace{Y, \mathbf{L}, \mathbf{A}}_{\text{data}} \,;\, \underbrace{\theta}_{\text{param.}})$$

that is **unbiased**, meaning

$$\mathrm{E}\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}; \underbrace{\theta_0}_{\text{true}})\} = \mathbf{0}.$$

The M-estimator $\widehat{\theta}$ is the solution to

Crash Course on M-Estimation

Score Function: a function of the observed data and the parameter of interest

$$\underbrace{\mathbf{\Psi}_0}_{\text{score fun.}} \left(\underbrace{Y, \mathbf{L}, \mathbf{A}}_{\text{data}} ; \underbrace{\theta}_{\text{param.}} \right)$$

that is **unbiased**, meaning

$$\mathrm{E}\{\mathbf{\Psi}_0(Y, \mathbf{L}, \mathbf{A}; \underbrace{\theta_0}_{\text{true}})\} = \mathbf{0}.$$

The M-estimator $\widehat{\theta}$ is the solution to

$$\textstyle\sum_{i=1}^{n} \mathbf{\Psi}_0(\underbrace{Y_i, \mathbf{L}_i, \mathbf{A}_i}_{\text{observed}}; \theta) = \mathbf{0}.$$

Crash Course on M-Estimation

Score Function: a function of the observed data and the parameter of interest

$$\underbrace{\boldsymbol{\Psi}_0}_{\text{score fun.}} \; (\underbrace{Y, \mathbf{L}, \mathbf{A}}_{\text{data}} \; ; \; \underbrace{\theta}_{\text{param.}} \;)$$

that is **unbiased**, meaning

$$\mathrm{E}\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}; \underbrace{\theta_0}_{\text{true}})\} = \mathbf{0}.$$

The M-estimator $\widehat{\theta}$ is the solution to

$$\textstyle\sum_{i=1}^{n} \boldsymbol{\Psi}_0(\underbrace{Y_i, \mathbf{L}_i, \mathbf{A}_i}_{\text{observed}}; \theta) = \mathbf{0}.$$

$\widehat{\theta}$ is consistent and asymptotically normal and has a simple variance estimator

G-Formula

1. Fit the **outcome model**

$$\mu(\mathbf{L}, \mathbf{A}; \beta) \equiv \mathrm{E}(Y|\mathbf{L}, \mathbf{A})$$

G-Formula

1. Fit the **outcome model**

$$\mu(\mathbf{L}, \mathbf{A}; \beta) \equiv \mathrm{E}(Y|\mathbf{L}, \mathbf{A})$$

2. Estimate the dose-response surface by marginalizing over the distribution of confounders:

$$\widehat{\eta}(\mathbf{a}) = n^{-1} \sum_{i=1}^{n} \mu(\mathbf{L}_i, \mathbf{a}; \widehat{\beta})$$

G-Formula

1. Fit the **outcome model**

$$\mu(\mathbf{L}, \mathbf{A}; \beta) \equiv \mathrm{E}(Y|\mathbf{L}, \mathbf{A})$$

2. Estimate the dose-response surface by marginalizing over the distribution of confounders:

$$\widehat{\eta}(\mathbf{a}) = n^{-1} \sum_{i=1}^{n} \mu(\mathbf{L}_i, \mathbf{a}; \widehat{\beta})$$

This can be expressed as an M-estimator with score function

$$\boldsymbol{\Psi}_{0-GF}(Y, \mathbf{L}, \mathbf{A}; \theta_{GF}) = \begin{bmatrix} \{Y - \mu(\mathbf{L}, \mathbf{A}; \beta)\} \, \partial_\beta \mu(\mathbf{L}, \mathbf{A}; \beta) \\ \eta(\mathbf{a}) - \mu(\mathbf{L}, \mathbf{a}; \beta) \end{bmatrix}$$

Inverse Probability Weighting (IPW)

1. Obtain/estimate **standardized propensity score weights**

$$SW(\mathbf{L}, \mathbf{A}) = \frac{f_{\mathbf{A}}(\mathbf{A})}{f_{\mathbf{A}|\mathbf{L}}(\mathbf{A}|\mathbf{L})}$$

Inverse Probability Weighting (IPW)

1. Obtain/estimate **standardized propensity score weights**

$$SW(\mathbf{L}, \mathbf{A}) = \frac{f_{\mathbf{A}}(\mathbf{A})}{f_{\mathbf{A}|\mathbf{L}}(\mathbf{A}|\mathbf{L})}$$

2. Use weighted observations to estimate the dose-response surface $\eta(\mathbf{a}; \gamma)$

Inverse Probability Weighting (IPW)

1. Obtain/estimate **standardized propensity score weights**

$$SW(\mathbf{L}, \mathbf{A}) = \frac{f_{\mathbf{A}}(\mathbf{A})}{f_{\mathbf{A}|\mathbf{L}}(\mathbf{A}|\mathbf{L})}$$

2. Use weighted observations to estimate the dose-response surface $\eta(\mathbf{a}; \gamma)$

This can be expressed as an M-estimator with score function

$$\boldsymbol{\Psi}_{0-IPW}(Y, \mathbf{L}, \mathbf{A}; \theta_{IPW}) = \begin{bmatrix} \boldsymbol{\Psi}_{PS}(\mathbf{L}, \mathbf{A}) \\ SW(\mathbf{L}, \mathbf{A}) \left\{ Y - \eta(\mathbf{A}; \gamma) \right\} \partial_\gamma \eta(\mathbf{A}; \gamma) \end{bmatrix}$$

Doubly Robust (DR)

1. Obtain/estimate **standardized propensity score weights** $SW(\mathbf{L}, \mathbf{A})$

Doubly Robust (DR)

1. Obtain/estimate **standardized propensity score weights** $SW(\mathbf{L}, \mathbf{A})$
2. Use weighted observations to estimate the **outcome model** $\mu(\mathbf{L}, \mathbf{A}; \beta) \equiv \mathrm{E}(Y|\mathbf{L}, \mathbf{A})$

Doubly Robust (DR)

1. Obtain/estimate **standardized propensity score weights** $SW(\mathbf{L}, \mathbf{A})$
2. Use weighted observations to estimate the **outcome model** $\mu(\mathbf{L}, \mathbf{A}; \beta) \equiv \mathrm{E}(Y|\mathbf{L}, \mathbf{A})$
3. Estimate the dose-response surface by marginalizing over the distribution of confounders

Doubly Robust (DR)

1. Obtain/estimate **standardized propensity score weights** $SW(\mathbf{L}, \mathbf{A})$
2. Use weighted observations to estimate the **outcome model** $\mu(\mathbf{L}, \mathbf{A}; \beta) \equiv \mathrm{E}(Y|\mathbf{L}, \mathbf{A})$
3. Estimate the dose-response surface by marginalizing over the distribution of confounders

This can be expressed as an M-estimator with score function

$$\boldsymbol{\Psi}_{0-DR}(Y, \mathbf{L}, \mathbf{A}; \theta_{DR}) = \begin{bmatrix} \boldsymbol{\Psi}_{PS}(\mathbf{L}, \mathbf{A}) \\ SW(\mathbf{L}, \mathbf{A})\{Y - \mu(\mathbf{L}, \mathbf{A}; \beta)\}\partial_\beta \mu(\mathbf{L}, \mathbf{A}; \beta) \\ \eta(\mathbf{a}) - \mu(\mathbf{L}, \mathbf{a}; \beta) \end{bmatrix}$$

Doubly Robust (DR)

1. Obtain/estimate **standardized propensity score weights** $SW(\mathbf{L}, \mathbf{A})$
2. Use weighted observations to estimate the **outcome model** $\mu(\mathbf{L}, \mathbf{A}; \beta) \equiv \mathrm{E}(Y|\mathbf{L}, \mathbf{A})$
3. Estimate the dose-response surface by marginalizing over the distribution of confounders

This can be expressed as an M-estimator with score function

$$\boldsymbol{\Psi}_{0-DR}(Y, \mathbf{L}, \mathbf{A}; \theta_{DR}) = \begin{bmatrix} \boldsymbol{\Psi}_{PS}(\mathbf{L}, \mathbf{A}) \\ SW(\mathbf{L}, \mathbf{A})\{Y - \mu(\mathbf{L}, \mathbf{A}; \beta)\}\partial_\beta \mu(\mathbf{L}, \mathbf{A}; \beta) \\ \eta(\mathbf{a}) - \mu(\mathbf{L}, \mathbf{a}; \beta) \end{bmatrix}$$

Doubly robust* to models for $\mu(\mathbf{L}, \mathbf{A}; \beta)$ and $f_{\mathbf{A}|\mathbf{L}}(\mathbf{A}|\mathbf{L})$

Addressing Confounding and Measurement Error

Can we just substitute $\mathbf{A}^*$ for $\mathbf{A}$ and find the solution to

$$\sum_{i=1}^{n} \mathbf{\Psi}_0(Y_i, \mathbf{L}_i, \underbrace{\mathbf{A}_i^*}_{\text{mismeasured}}; \theta) = \mathbf{0}?$$

Addressing Confounding and Measurement Error

Can we just substitute $\mathbf{A}^*$ for $\mathbf{A}$ and find the solution to

$$\sum_{i=1}^{n} \mathbf{\Psi}_0(Y_i, \mathbf{L}_i, \underbrace{\mathbf{A}_i^*}_{\text{mismeasured}}; \theta) = \mathbf{0}?$$

No! This leads to bias in $\widehat{\theta}$ because

$$\mathrm{E}\{\mathbf{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^*; \theta_0)\} \neq \mathbf{0}.$$

Addressing Confounding and Measurement Error

Can we just substitute $\mathbf{A}^*$ for $\mathbf{A}$ and find the solution to

$$\sum_{i=1}^n \mathbf{\Psi}_0(Y_i, \mathbf{L}_i, \underbrace{\mathbf{A}_i^*}_{\text{mismeasured}}; \theta) = \mathbf{0}?$$

No! This leads to bias in $\widehat{\theta}$ because

$$\mathrm{E}\{\mathbf{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^*; \theta_0)\} \neq \mathbf{0}.$$

We need a new score function $\mathbf{\Psi}_{CS}$ such that

$$\mathrm{E}\{\underbrace{\mathbf{\Psi}_{CS}}_{\text{new score fun.}}(Y, \mathbf{L}, \underbrace{\mathbf{A}^*}_{\text{mismeasured}}; \theta_0)\} = \mathbf{0}.$$

Corrected Score Functions

Given the "oracle" score function $\mathbf{\Psi}_0$, the "corrected score" function $\mathbf{\Psi}_{CS}$ can be created following Novick and Stefanski (2002) (Novick and Stefanski, 2002):

$$\mathbf{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^*; \theta)$$

Corrected Score Functions

Given the "oracle" score function $\mathbf{\Psi}_0$, the "corrected score" function $\mathbf{\Psi}_{CS}$ can be created following Novick and Stefanski (2002) (Novick and Stefanski, 2002):

$$\mathbf{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^* + i\widetilde{\epsilon}; \theta)$$

1. Add additional **imaginary** measurement error $i\widetilde{\epsilon}$ to the mismeasured exposure

Corrected Score Functions

Given the "oracle" score function $\mathbf{\Psi}_0$, the "corrected score" function $\mathbf{\Psi}_{CS}$ can be created following Novick and Stefanski (2002) (Novick and Stefanski, 2002):

$$\text{Re}\{\mathbf{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^* + i\widetilde{\epsilon}; \theta)\}$$

1. Add additional **imaginary** measurement error $i\widetilde{\epsilon}$ to the mismeasured exposure
2. Keep only the real part of the resulting complex-valued function

Corrected Score Functions

Given the "oracle" score function $\mathbf{\Psi}_0$, the "corrected score" function $\mathbf{\Psi}_{CS}$ can be created following Novick and Stefanski (2002) (Novick and Stefanski, 2002):

$$E\left[\mathrm{Re}\left\{\mathbf{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^* + i\widetilde{\epsilon}; \theta)\right\} \mid Y, \mathbf{L}, \mathbf{A}^*\right]$$

1. Add additional **imaginary** measurement error $i\widetilde{\epsilon}$ to the mismeasured exposure
2. Keep only the real part of the resulting complex-valued function
3. Take the expectation over the additional measurement error $\widetilde{\epsilon}$.

Corrected Score Functions

Given the "oracle" score function $\boldsymbol{\Psi}_0$, the "corrected score" function $\boldsymbol{\Psi}_{CS}$ can be created following Novick and Stefanski (2002) (Novick and Stefanski, 2002):

$$\boldsymbol{\Psi}_{CS}(Y, \mathbf{L}, \mathbf{A}^*; \theta) \equiv \mathrm{E}\left[\mathrm{Re}\left\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^* + i\widetilde{\epsilon}; \theta)\right\} \mid Y, \mathbf{L}, \mathbf{A}^*\right]$$

1. Add additional **imaginary** measurement error $i\widetilde{\epsilon}$ to the mismeasured exposure
2. Keep only the real part of the resulting complex-valued function
3. Take the expectation over the additional measurement error $\widetilde{\epsilon}$.

Corrected Score Functions (contd)

If the oracle score function is **conditionally unbiased**:

$$\mathrm{E}\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}; \theta_0) | \mathbf{A}\} = \mathbf{0}$$

Corrected Score Functions (contd)

If the oracle score function is **conditionally unbiased**:

$$\mathrm{E}\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}; \theta_0)|\mathbf{A}\} = \mathbf{0}$$

then the corrected score function $\boldsymbol{\Psi}_{CS}$ is unbiased, meaning

$$\mathrm{E}\{\boldsymbol{\Psi}_{CS}(Y, \mathbf{L}, \underbrace{\mathbf{A}^*}_{\text{mismeasured}}; \theta_0)\} = \mathbf{0}$$

Corrected Score Functions (contd)

If the oracle score function is **conditionally unbiased**:

$$\mathrm{E}\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}; \theta_0) | \mathbf{A}\} = \mathbf{0}$$

then the corrected score function $\boldsymbol{\Psi}_{CS}$ is unbiased, meaning

$$\mathrm{E}\{\boldsymbol{\Psi}_{CS}(Y, \mathbf{L}, \underbrace{\mathbf{A}^*}_{\text{mismeasured}}; \theta_0)\} = \mathbf{0}$$

The G-Formula, IPW, and DR score functions all satisfy these conditions, and so can be "corrected:"

$$\boldsymbol{\Psi}_{0-GF} \longrightarrow \boldsymbol{\Psi}_{CS-GF}$$
$$\boldsymbol{\Psi}_{0-IPW} \longrightarrow \boldsymbol{\Psi}_{CS-IPW}$$
$$\boldsymbol{\Psi}_{0-DR} \longrightarrow \boldsymbol{\Psi}_{CS-DR}$$

Monte Carlo Corrected Score Functions

- Sometimes we can find a closed-form algebraic expression for

$$\boldsymbol{\Psi}_{CS}(Y, \mathbf{L}, \mathbf{A}^*; \theta) = \underbrace{\mathrm{E}\left[\mathrm{Re}\left\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^* + i\widetilde{\epsilon}; \theta)\right\} \middle| Y, \mathbf{L}, \mathbf{A}^*\right]}_{\mathrm{E}\{f(\widetilde{\epsilon})\}}.$$

Monte Carlo Corrected Score Functions

- Sometimes we can find a closed-form algebraic expression for

$$\boldsymbol{\Psi}_{CS}\left(Y, \mathbf{L}, \mathbf{A}^*; \theta\right) = \underbrace{\mathrm{E}\left[\mathrm{Re}\left\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^* + i\widetilde{\epsilon}; \theta)\right\} \middle| Y, \mathbf{L}, \mathbf{A}^*\right]}_{\mathrm{E}\{f(\widetilde{\epsilon})\}}.$$

- Alternatively, we can approximate this expectation with Monte Carlo replicates

Monte Carlo Corrected Score Functions

- Sometimes we can find a closed-form algebraic expression for

$$\mathbf{\Psi}_{CS}\left(Y, \mathbf{L}, \mathbf{A}^*; \theta\right) = \underbrace{\mathrm{E}\left[\mathrm{Re}\left\{\mathbf{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^* + i\widetilde{\epsilon}; \theta)\right\} \middle| Y, \mathbf{L}, \mathbf{A}^*\right]}_{\mathrm{E}\{f(\widetilde{\epsilon})\}}.$$

- Alternatively, we can approximate this expectation with Monte Carlo replicates

$$\mathrm{E}\{f(\widetilde{\epsilon})\} \approx B^{-1} \textstyle\sum_{b=1}^{B} f(\widetilde{\epsilon}_b)$$

Monte Carlo Corrected Score Functions

- Sometimes we can find a closed-form algebraic expression for

$$\boldsymbol{\Psi}_{CS}\left(Y, \mathbf{L}, \mathbf{A}^*; \theta\right) = \underbrace{\mathrm{E}\left[\mathrm{Re}\left\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^* + i\widetilde{\epsilon}; \theta)\right\} \mid Y, \mathbf{L}, \mathbf{A}^*\right]}_{\mathrm{E}\{f(\widetilde{\epsilon})\}}.$$

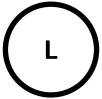
- Alternatively, we can approximate this expectation with Monte Carlo replicates

$$\mathrm{E}\{f(\widetilde{\epsilon})\} \approx B^{-1} \textstyle\sum_{b=1}^{B} f(\widetilde{\epsilon}_b)$$

$$\implies \boldsymbol{\Psi}_{MCCS}^{B}(Y, \mathbf{L}, \mathbf{A}^*; \theta) = B^{-1} \textstyle\sum_{b=1}^{B} \mathrm{Re}\left\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}^* + i\widetilde{\epsilon}_b; \theta)\right\}$$

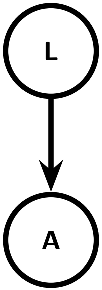
Simulation Setting

$$L \sim \mathcal{N}(0, 0.36)$$



Simulation Setting

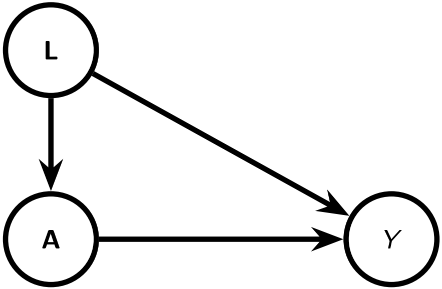
$L \sim \mathcal{N}(0, 0.36)$



$\mathbf{A}|L \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{I})$

Simulation Setting

$L \sim \mathcal{N}(0, 0.36)$



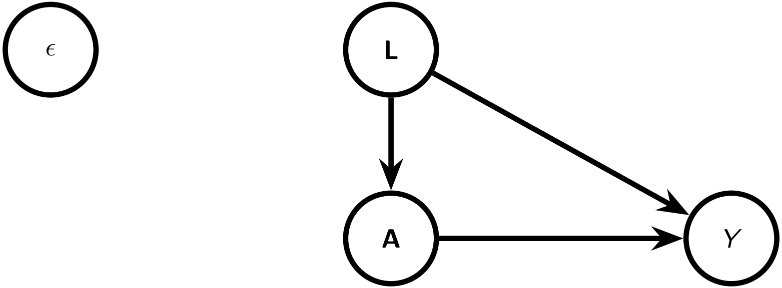
$\mathbf{A} | L \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{I})$

$Y | L, \mathbf{A} \sim \mathcal{N}(A_1 + A_2 + L, 1)$

Simulation Setting

$\epsilon \sim \mathcal{N}_2(\mathbf{0}, \sigma^2_{me}\boldsymbol{I})$

$L \sim \mathcal{N}(0, 0.36)$



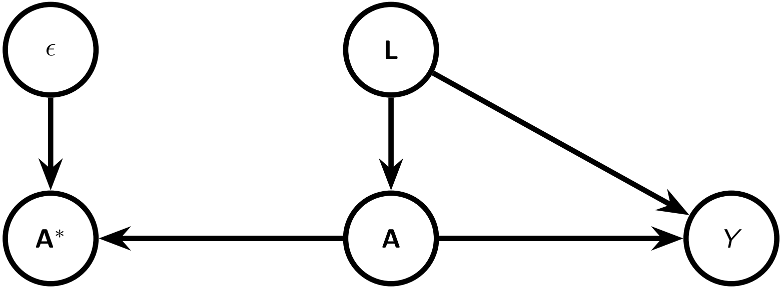
$\mathbf{A}|L \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{I})$

$Y|L, \mathbf{A} \sim \mathcal{N}(A_1 + A_2 + L, 1)$

Simulation Setting

$\epsilon \sim \mathcal{N}_2(\mathbf{0}, \sigma^2_{me}\boldsymbol{I})$

$L \sim \mathcal{N}(0, 0.36)$

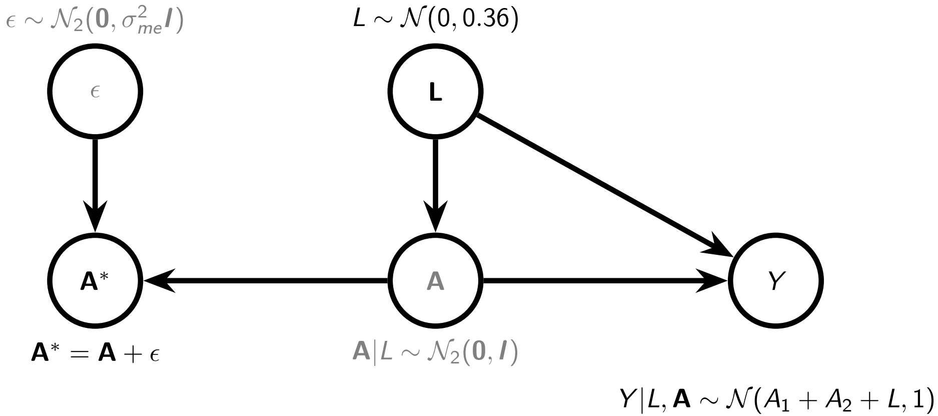


$\mathbf{A}^* = \mathbf{A} + \epsilon$

$\mathbf{A}|L \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{I})$

$Y|L, \mathbf{A} \sim \mathcal{N}(A_1 + A_2 + L, 1)$

Simulation Setting



Simulation Setting

- Implies a dose-response surface of $\eta(\mathbf{a};\gamma) = \gamma_0 + \underbrace{\gamma_1}_{\text{estimand}} a_1 + \gamma_2 a_2$

Simulation Setting

- Implies a dose-response surface of $\eta(\mathbf{a};\gamma) = \gamma_0 + \underbrace{\gamma_1}_{\text{estimand}} a_1 + \gamma_2 a_2$
- Sample size $n = 800$

Simulation Setting

- Implies a dose-response surface of $\eta(\mathbf{a};\gamma) = \gamma_0 + \underbrace{\gamma_1}_{\text{estimand}} a_1 + \gamma_2 a_2$
- Sample size n $= 800$
- 2 estimators compared:

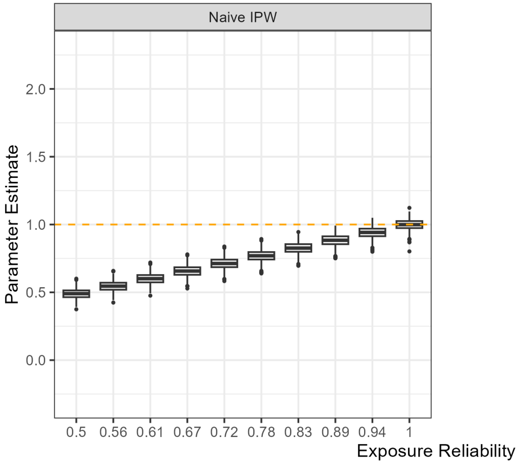
Simulation Setting

- Implies a dose-response surface of $\eta(\mathbf{a};\gamma) = \gamma_0 + \underbrace{\gamma_1}_{\text{estimand}} a_1 + \gamma_2 a_2$
- Sample size n $=$ 800
- 2 estimators compared:
  1. Naive IPW (ignores measurement error)

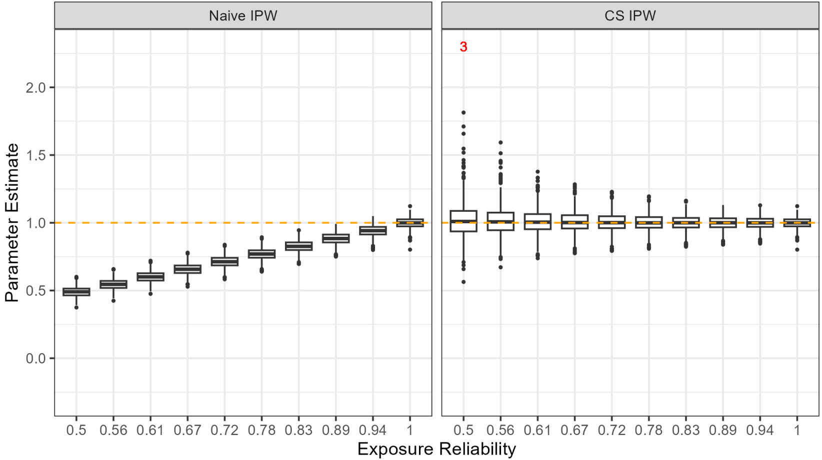
Simulation Setting

- Implies a dose-response surface of $\eta(\mathbf{a};\gamma) = \gamma_0 + \underbrace{\gamma_1}_{\text{estimand}} a_1 + \gamma_2 a_2$
- Sample size n $= 800$
- 2 estimators compared:
  1. Naive IPW (ignores measurement error)
  2. Corrected Score IPW

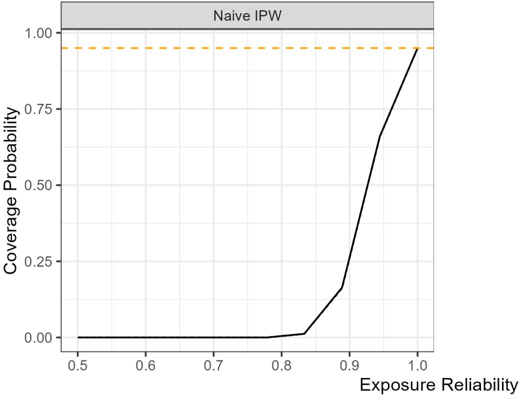
Simulation Results: Estimator



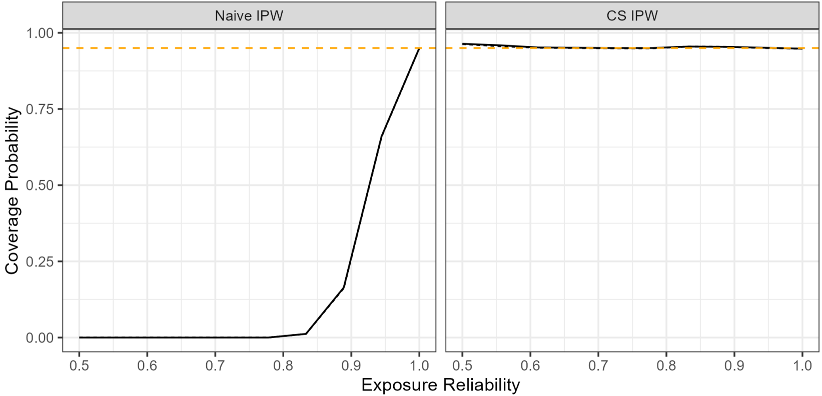
Simulation Results: Estimator



Simulation Results: Confidence Interval



Simulation Results: Confidence Interval



Application: HVTN 505 Trial

- **Two exposures** (both log-transformed):

Application: HVTN 505 Trial

- **Two exposures** (both log-transformed):
  (i) Antibody-dependent cellular phagocytosis (**ADCP**)

Application: HVTN 505 Trial

- **Two exposures** (both log-transformed):
  (i) Antibody-dependent cellular phagocytosis (**ADCP**)
  (ii) Recruitment of FcγRIIa of the H131-Con S gp140 protein (**RII**)

Application: HVTN 505 Trial

- **Two exposures** (both log-transformed):
  (i) Antibody-dependent cellular phagocytosis (**ADCP**)
  (ii) Recruitment of FcγRIIa of the H131-Con S gp140 protein (**RII**)
- **Case-cohort sampling**: biomarkers only measured in stratified random sample of controls

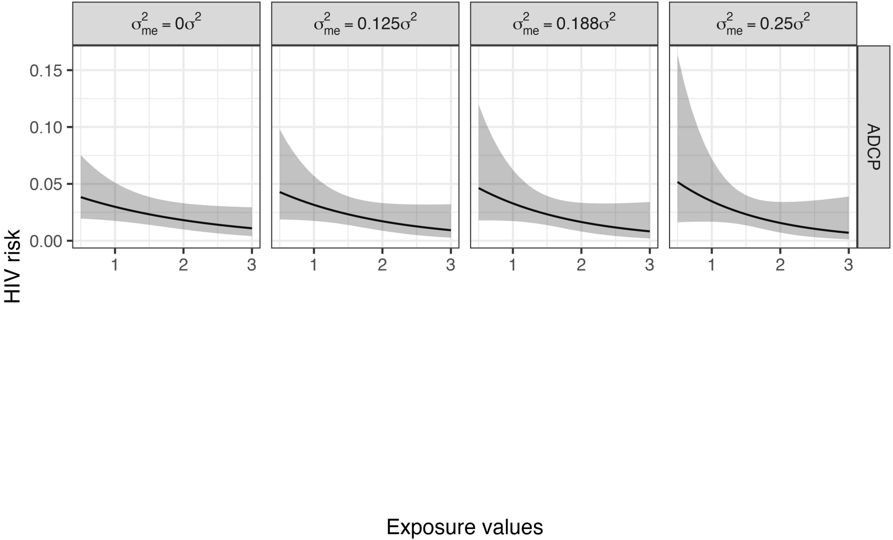
Application: HVTN 505 Trial

- **Two exposures** (both log-transformed):
  (i) Antibody-dependent cellular phagocytosis (**ADCP**)
  (ii) Recruitment of FcγRIIa of the H131-Con S gp140 protein (**RII**)
- **Case-cohort sampling**: biomarkers only measured in stratified random sample of controls
- **Covariates**: age, race, BMI, behavior risk, CD4-P, and CD8-P

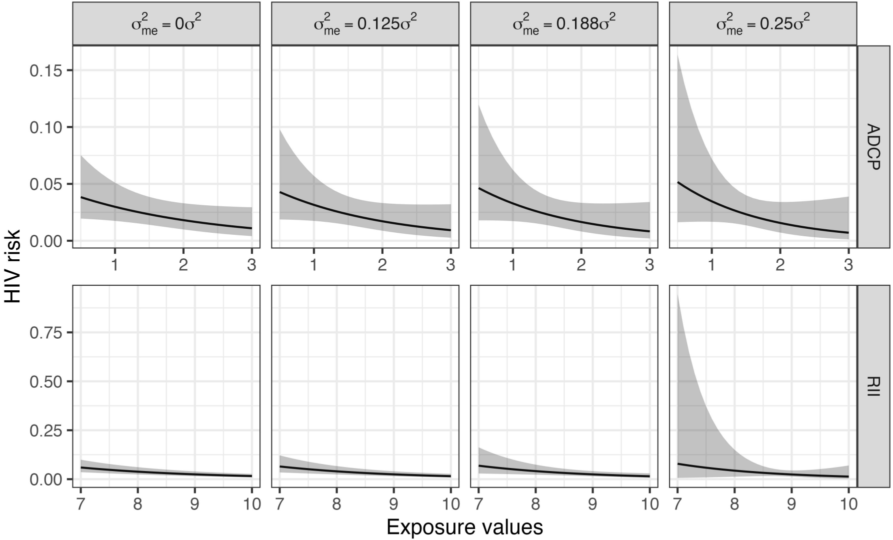
Application: HVTN 505 Trial

- **Two exposures** (both log-transformed):
  (i) Antibody-dependent cellular phagocytosis (**ADCP**)
  (ii) Recruitment of FcγRIIa of the H131-Con S gp140 protein (**RII**)
- **Case-cohort sampling**: biomarkers only measured in stratified random sample of controls
- **Covariates**: age, race, BMI, behavior risk, CD4-P, and CD8-P
- **Analysis**: Corrected score DR estimator with a log-linear outcome model

Application: HVTN 505 Trial



Application: HVTN 505 Trial



Summary of Project 1

Paper in *Biometrics* (2025)

GitHub R package

Mismex: Causal Inference for **Mism**easured **Ex**posures

Project 2: Social Distancing to Reduce Transmission of Influenza-Like-Illness on College Campuses: the eX-FLU Trial

Brian Richardson, Allison Aiello, Michael Hudgens

eX-FLU Trial

- **eX-FLU**: trial to evaluate a social distancing intervention on a college campus during flu season (Aiello et al., 2016; Zivich et al., 2020)

Design and methods of a social network isolation study for reducing respiratory infection transmission: The eX-FLU cluster randomized trial

Allison E. Aiello[a,*], Amanda M. Simanek[b,1], Marisa C. Eisenberg[c], Alison R. Walsh[c], Brian Davis[c], Erik Volz[d,1], Caroline Cheng[c], Jeanette J. Rainey[e], Amra Uzicanin[e], Hongjiang Gao[e], Nathaniel Osgood[f], Dylan Knowles[f], Kevin Stanley[f], Kara Tarter[c], Arnold S. Monto[c]

eX-FLU Trial

- **eX-FLU**: trial to evaluate a social distancing intervention on a college campus during flu season (Aiello et al., 2016; Zivich et al., 2020)
- **Intervention**: encouragement to isolate in dorm for three days upon developing symptoms of **influenza-like illness** (ILI)

Design and methods of a social network isolation study for reducing respiratory infection transmission: The eX-FLU cluster randomized trial

Allison E. Aiello[a,*], Amanda M. Simanek[b,1], Marisa C. Eisenberg[c], Alison R. Walsh[c], Brian Davis[c], Erik Volz[d,1], Caroline Cheng[c], Jeanette J. Rainey[e], Amra Uzicanin[e], Hongjiang Gao[e], Nathaniel Osgood[f], Dylan Knowles[f], Kevin Stanley[f], Kara Tarter[c], Arnold S. Monto[c]

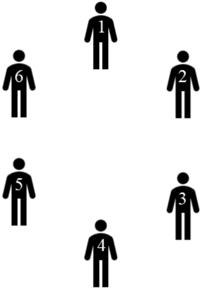
eX-FLU Trial

- **eX-FLU**: trial to evaluate a social distancing intervention on a college campus during flu season (Aiello et al., 2016; Zivich et al., 2020)
- **Intervention**: encouragement to isolate in dorm for three days upon developing symptoms of **influenza-like illness** (ILI)
- **Central question**: does the encouragement-to-isolate intervention reduce transmission of ILI?

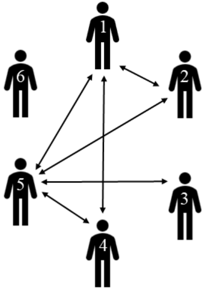
Design and methods of a social network isolation study for reducing respiratory infection transmission: The eX-FLU cluster randomized trial

Allison E. Aiello[a,*], Amanda M. Simanek[b,1], Marisa C. Eisenberg[c], Alison R. Walsh[c], Brian Davis[c], Erik Volz[d,1], Caroline Cheng[c], Jeanette J. Rainey[e], Amra Uzicanin[e], Hongjiang Gao[e], Nathaniel Osgood[f], Dylan Knowles[f], Kevin Stanley[f], Kara Tarter[c], Arnold S. Monto[c]

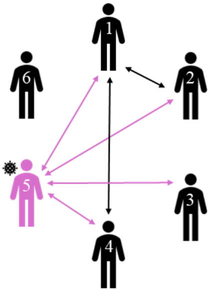
Transmission of Influenza-Like-Illness



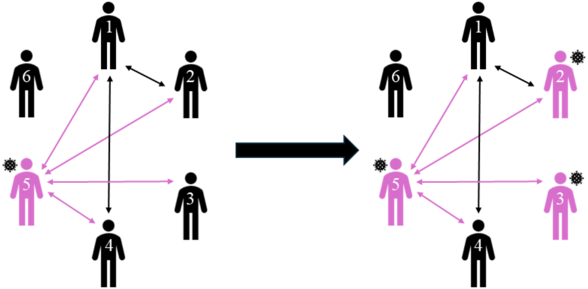
Transmission of Influenza-Like-Illness



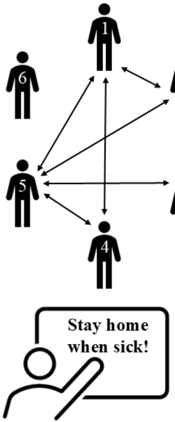
Transmission of Influenza-Like-Illness



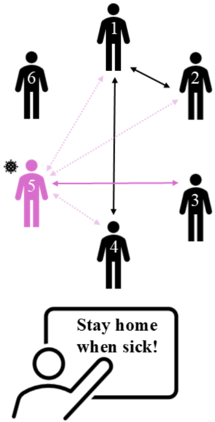
Transmission of Influenza-Like-Illness



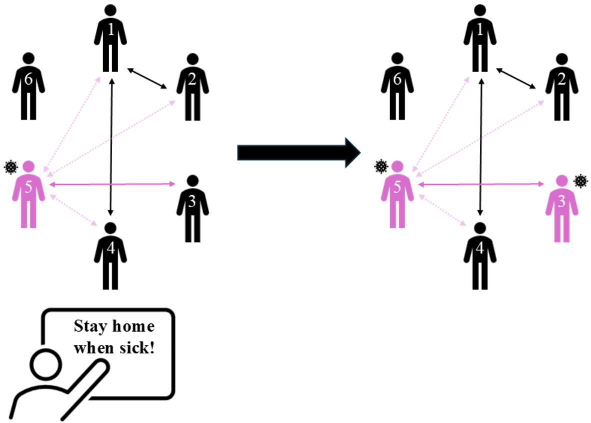
Example I: Intervention Affects Network and Transmission



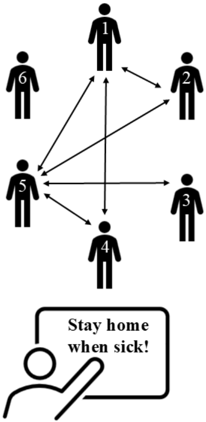
Example I: Intervention Affects Network and Transmission



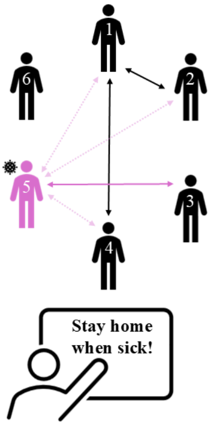
Example I: Intervention Affects Network and Transmission



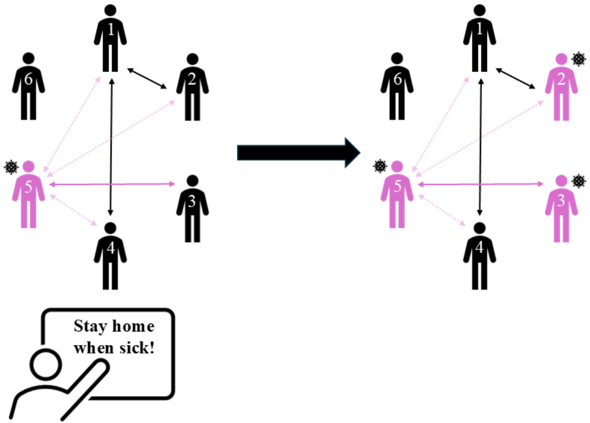
Example II: Intervention Affects Network, Not Transmission



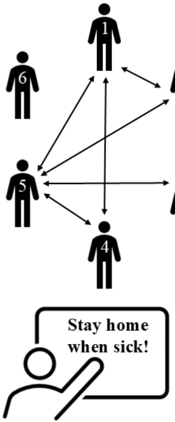
Example II: Intervention Affects Network, Not Transmission



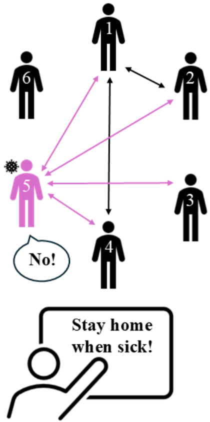
Example II: Intervention Affects Network, Not Transmission



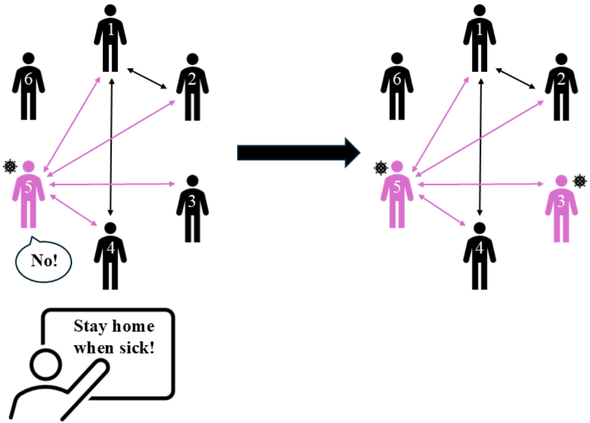
Example III: Intervention Affects Transmission, Not Network



Example III: Intervention Affects Transmission, Not Network

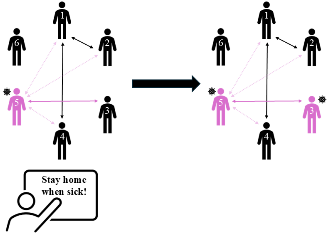


Example III: Intervention Affects Transmission, Not Network

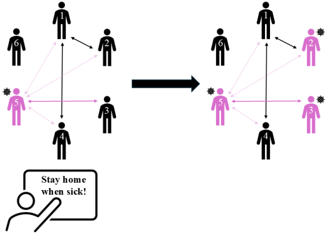


Central Question: Does the Intervention Affect Transmission of ILI?

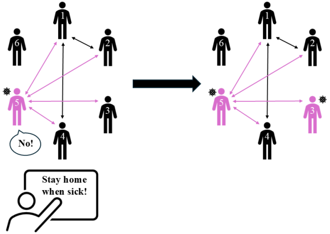
Example I:



Example II:



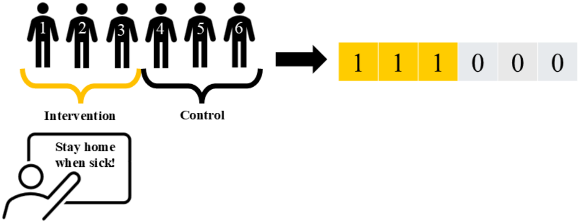
Example III:



In Examples I and III, the answer is yes

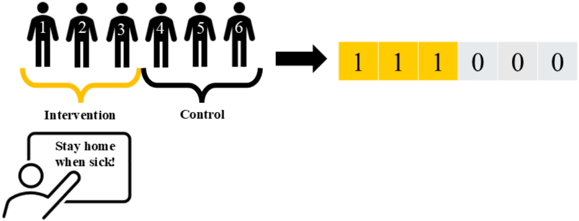
eX-FLU Observed Data

- Baseline randomization assignments
  $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathcal{Z}$, for
  $Z_i = \mathbb{1}(\text{student } i \text{ gets intervention})$



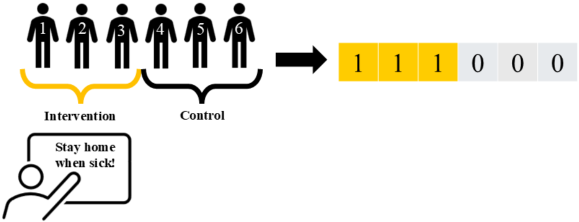
eX-FLU Observed Data

- Baseline randomization assignments $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathcal{Z}$, for $Z_i = \mathbb{1}(\text{student } i \text{ gets intervention})$
    - known randomization distribution $r(\mathbf{z})$



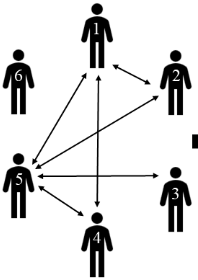
eX-FLU Observed Data

- Baseline randomization assignments $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathcal{Z}$, for $Z_i = \mathbb{1}(\text{student } i \text{ gets intervention})$
  - known randomization distribution $r(\mathbf{z})$
- At weeks $k = 1, \ldots, \tau$, we observe:



eX-FLU Observed Data

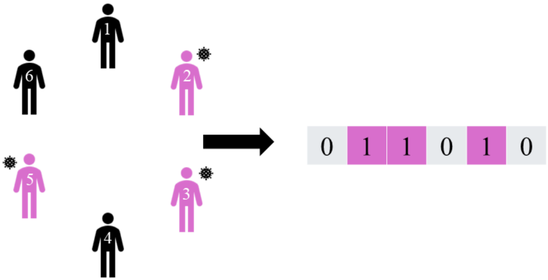
- Baseline randomization assignments
  $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathcal{Z}$, for
  $Z_i = \mathbb{1}(\text{student } i \text{ gets intervention})$
  - known randomization distribution $r(\mathbf{z})$
- At weeks $k = 1, \ldots, \tau$, we observe:
  - networks $\mathbf{A}^k = [A^k_{ij}]$, for
    $A^k_{ij} = \mathbb{1}(\text{students } i, j \text{ in contact at week } k)$



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

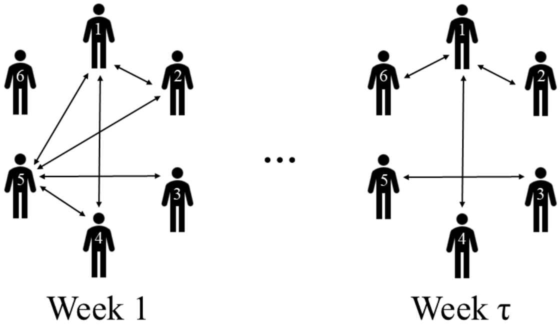
eX-FLU Observed Data

- Baseline randomization assignments $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathcal{Z}$, for $Z_i = \mathbb{1}(\text{student } i \text{ gets intervention})$
  - known randomization distribution $r(\mathbf{z})$
- At weeks $k = 1, \ldots, \tau$, we observe:
  - networks $\mathbf{A}^k = [A^k_{ij}]$, for $A^k_{ij} = \mathbb{1}(\text{students } i, j \text{ in contact at week } k)$
  - infections $\mathbf{Y}^k = (Y^k_1, \ldots, Y^k_n)$, for $Y^k_i = \mathbb{1}(\text{student } i \text{ is infected at week } k)$



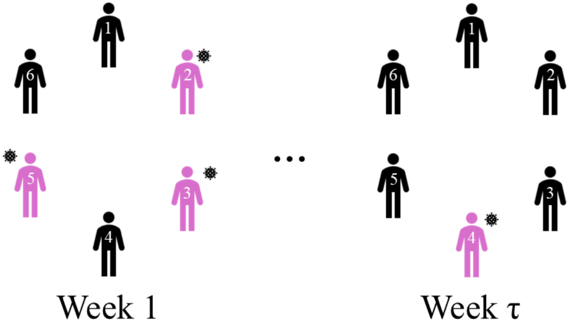
eX-FLU Observed Data

- Baseline randomization assignments $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathcal{Z}$, for $Z_i = \mathbb{1}(\text{student } i \text{ gets intervention})$
    - known randomization distribution $r(\mathbf{z})$
- At weeks $k = 1, \ldots, \tau$, we observe:
    - networks $\mathbf{A}^k = [A_{ij}^k]$, for $A_{ij}^k = \mathbb{1}(\text{students } i, j \text{ in contact at week } k)$
    - infections $\mathbf{Y}^k = (Y_1^k, \ldots, Y_n^k)$, for $Y_i^k = \mathbb{1}(\text{student } i \text{ is infected at week } k)$
- network history: $\overline{\mathbf{A}} = \{\mathbf{A}^k\}_{k=1}^{\tau}$



eX-FLU Observed Data

- Baseline randomization assignments $\mathbf{Z} = (Z_1, \ldots, Z_n) \in \mathcal{Z}$, for $Z_i = \mathbb{1}(\text{student } i \text{ gets intervention})$
  - known randomization distribution $r(\mathbf{z})$
- At weeks $k = 1, \ldots, \tau$, we observe:
  - networks $\mathbf{A}^k = [A^k_{ij}]$, for $A^k_{ij} = \mathbb{1}(\text{students } i, j \text{ in contact at week } k)$
  - infections $\mathbf{Y}^k = (Y^k_1, \ldots, Y^k_n)$, for $Y^k_i = \mathbb{1}(\text{student } i \text{ is infected at week } k)$
- network history: $\overline{\mathbf{A}} = \{\mathbf{A}^k\}_{k=1}^{\tau}$
- infection history: $\overline{\mathbf{Y}} = \{\mathbf{Y}^k\}_{k=1}^{\tau}$



eX-FLU Potential Outcomes

Potential outcomes are defined for both the networks, and ILI infections:

$$\overline{\mathbf{A}}(\mathbf{z}), \quad \overline{\mathbf{Y}}(\mathbf{z}) \quad \text{for } \mathbf{z} \in \mathcal{Z}$$

- $\overline{\mathbf{A}}(\mathbf{z})$: sequence of networks that **would** occur if students **were** encouraged to isolate according to $\mathbf{z}$
- $\overline{\mathbf{Y}}(\mathbf{z})$: sequence of infection statuses that **would** occur if students **were** encouraged to isolate according to $\mathbf{z}$

eX-FLU Potential Outcomes

Potential outcomes are defined for both the networks, and ILI infections:

$$\overline{\mathbf{A}}(\mathbf{z}), \quad \overline{\mathbf{Y}}(\mathbf{z}) \quad \text{for } \mathbf{z} \in \mathcal{Z}$$

- $\overline{\mathbf{A}}(\mathbf{z})$: sequence of networks that **would** occur if students **were** encouraged to isolate according to $\mathbf{z}$
- $\overline{\mathbf{Y}}(\mathbf{z})$: sequence of infection statuses that **would** occur if students **were** encouraged to isolate according to $\mathbf{z}$
- Both of these depend on the entire vector $\mathbf{z}$
  - potential outcomes for one student depend on intervention assignment of other students (**interference**) (Rubin, 1980; Sobel, 2006; Rosenbaum, 2007; Hudgens and Halloran, 2008)

eX-FLU Potential Outcomes

Potential outcomes are defined for both the networks, and ILI infections:

$$\overline{\mathbf{A}}(\mathbf{z}), \quad \overline{\mathbf{Y}}(\mathbf{z}) \quad \text{for } \mathbf{z} \in \mathcal{Z}$$

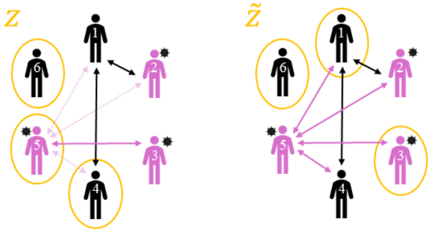
- $\overline{\mathbf{A}}(\mathbf{z})$: sequence of networks that **would** occur if students **were** encouraged to isolate according to $\mathbf{z}$
- $\overline{\mathbf{Y}}(\mathbf{z})$: sequence of infection statuses that **would** occur if students **were** encouraged to isolate according to $\mathbf{z}$
- Both of these depend on the entire vector $\mathbf{z}$
  - potential outcomes for one student depend on intervention assignment of other students (**interference**) (Rubin, 1980; Sobel, 2006; Rosenbaum, 2007; Hudgens and Halloran, 2008)
- Assume **causal consistency**:

$$\overline{\mathbf{A}} = \overline{\mathbf{A}}(\mathbf{Z}) \text{ and } \overline{\mathbf{Y}} = \overline{\mathbf{Y}}(\mathbf{Z})$$

eX-FLU Null Hypotheses

$$H_0^Y : \overline{\mathbf{Y}}(\mathbf{z}) = \overline{\mathbf{Y}}(\widetilde{\mathbf{z}}) \text{ for all } \mathbf{z}, \widetilde{\mathbf{z}} \in \mathcal{Z}$$

(“the intervention has no effect on the infections”)



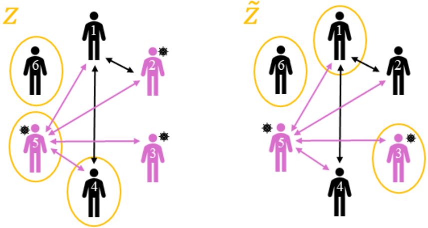
eX-FLU Null Hypotheses

$$H_0^Y : \overline{\mathbf{Y}}(\mathbf{z}) = \overline{\mathbf{Y}}(\widetilde{\mathbf{z}}) \text{ for all } \mathbf{z}, \widetilde{\mathbf{z}} \in \mathcal{Z}$$

("the intervention has no effect on the infections")

$$H_0^A : \overline{\mathbf{A}}(\mathbf{z}) = \overline{\mathbf{A}}(\widetilde{\mathbf{z}}) \text{ for all } \mathbf{z}, \widetilde{\mathbf{z}} \in \mathcal{Z}$$

("the intervention has no effect the networks")



eX-FLU Null Hypotheses

$$H_0^Y : \overline{\mathbf{Y}}(\mathbf{z}) = \overline{\mathbf{Y}}(\tilde{\mathbf{z}}) \text{ for all } \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$$

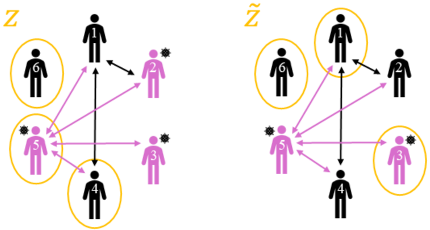
("the intervention has no effect on the infections")

$$H_0^A : \overline{\mathbf{A}}(\mathbf{z}) = \overline{\mathbf{A}}(\tilde{\mathbf{z}}) \text{ for all } \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$$

("the intervention has no effect the networks")

$$H_0^\sharp = H_0^Y \cap H_0^A : \overline{\mathbf{Y}}(\mathbf{z}) = \overline{\mathbf{Y}}(\tilde{\mathbf{z}}) \text{ and } \overline{\mathbf{A}}(\mathbf{z}) = \overline{\mathbf{A}}(\tilde{\mathbf{z}}) \text{ for all } \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$$

("the intervention has no effect on the infections or on the networks")



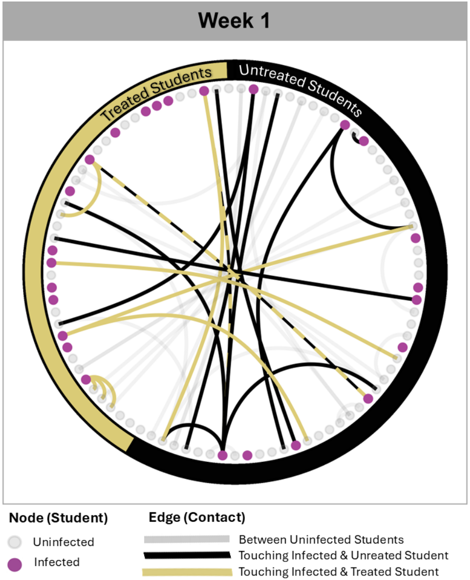
eX-FLU Null Hypotheses

$$H_0^Y : \overline{\mathbf{Y}}(\mathbf{z}) = \overline{\mathbf{Y}}(\widetilde{\mathbf{z}}) \text{ for all } \mathbf{z}, \widetilde{\mathbf{z}} \in \mathcal{Z}$$
$$(\text{“the intervention has no effect on the infections”})$$
$$H_0^A : \overline{\mathbf{A}}(\mathbf{z}) = \overline{\mathbf{A}}(\widetilde{\mathbf{z}}) \text{ for all } \mathbf{z}, \widetilde{\mathbf{z}} \in \mathcal{Z}$$
$$(\text{“the intervention has no effect the networks”})$$
$$H_0^\sharp = H_0^Y \cap H_0^A : \overline{\mathbf{Y}}(\mathbf{z}) = \overline{\mathbf{Y}}(\widetilde{\mathbf{z}}) \text{ and } \overline{\mathbf{A}}(\mathbf{z}) = \overline{\mathbf{A}}(\widetilde{\mathbf{z}}) \text{ for all } \mathbf{z}, \widetilde{\mathbf{z}} \in \mathcal{Z}$$
$$(\text{“the intervention has no effect on the infections or on the networks”})$$

Previous analyses of the eX-FLU trial (Alexandria et al., 2023) tested $H_0^\sharp$, but no analyses have tested H_0^Y

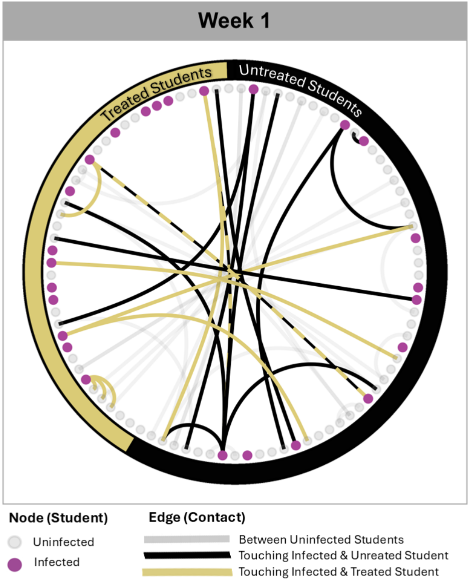
eX-FLU: Descriptive Results

- 93 (out of 579) students with at least one infection



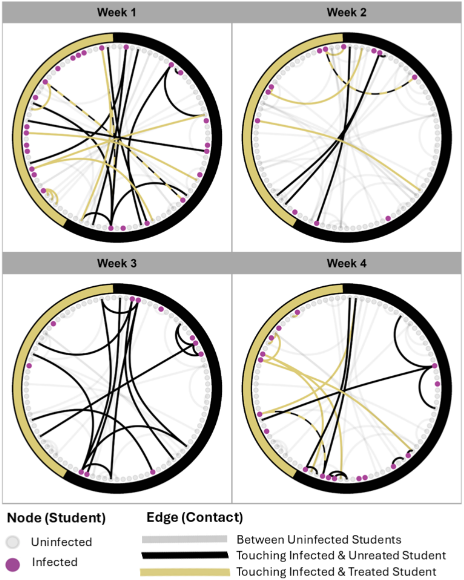
eX-FLU: Descriptive Results

- 93 (out of 579) students with at least one infection
- Bold edges represent possible transmission events



eX-FLU: Descriptive Results

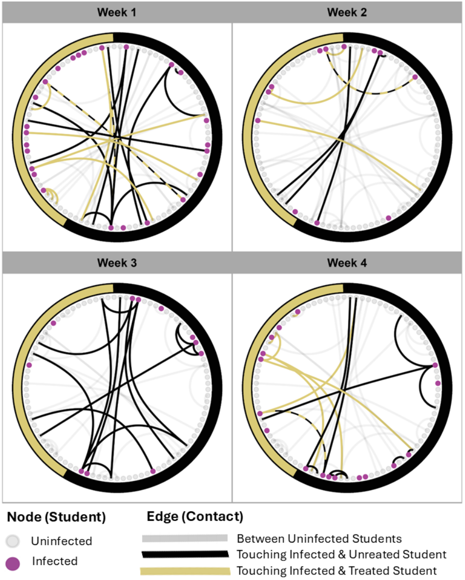
- 93 (out of 579) students with at least one infection
- Bold edges represent possible transmission events



eX-FLU: Descriptive Results

- 93 (out of 579) students with at least one infection
- Bold edges represent possible transmission events
  - Proportion of possible transmission events attributable to students in the intervention group:

$$
\begin{aligned}
T &= \frac{\text{number of yellow edges}}{\text{total number of edges}} = \\
&= \frac{\sum_{k=2}^{\tau}\sum_{i=1}^{n}\sum_{j>i} Z_i Y_i^{k-1} A_{ij}^{k-1}}{\sum_{k=2}^{\tau}\sum_{i=1}^{n}\sum_{j>i} Y_i^{k-1} A_{ij}^{k-1}} \\
&= 0.359
\end{aligned}
$$



Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?"

Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?" We can answer this using a **randomization test** (Fisher, 1936; Alexandria et al., 2023; Zhang and Zhao, 2023; Ritzwoller et al., 2024)

Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?" We can answer this using a **randomization test** (Fisher, 1936; Alexandria et al., 2023; Zhang and Zhao, 2023; Ritzwoller et al., 2024)

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$

Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?" We can answer this using a **randomization test** (Fisher, 1936; Alexandria et al., 2023; Zhang and Zhao, 2023; Ritzwoller et al., 2024)

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under $H_0^\sharp$

Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?" We can answer this using a **randomization test** (Fisher, 1936; Alexandria et al., 2023; Zhang and Zhao, 2023; Ritzwoller et al., 2024)

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under $H_0^\sharp$

Randomization	Networks	Infections	Test Statistic
$\mathbf{z}_1$			
$\mathbf{z}_2$			
$\ldots$			
$\mathbf{z}_{\lvert\mathcal{Z}\rvert}$			

Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?" We can answer this using a **randomization test** (Fisher, 1936; Alexandria et al., 2023; Zhang and Zhao, 2023; Ritzwoller et al., 2024)

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under $H_0^\sharp$

Randomization	Networks	Infections	Test Statistic
$\mathbf{z}_1$	$\overline{\mathbf{A}}(\mathbf{z}_1) = ?$	$\overline{\mathbf{Y}}(\mathbf{z}_1) = ?$	$T(\mathbf{z}_1, ?, ?)$
$\mathbf{z}_2$	$\overline{\mathbf{A}}(\mathbf{z}_2) = ?$	$\overline{\mathbf{Y}}(\mathbf{z}_1) = ?$	$T(\mathbf{z}_1, ?, ?)$
...	...	...	...
$\mathbf{z}_{\lvert\mathcal{Z}\rvert}$	$\overline{\mathbf{A}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = ?$	$\overline{\mathbf{Y}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = ?$	$T(\mathbf{z}_{\lvert\mathcal{Z}\rvert}, ?, ?)$

Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?" We can answer this using a **randomization test** (Fisher, 1936; Alexandria et al., 2023; Zhang and Zhao, 2023; Ritzwoller et al., 2024)

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under $H_0^\sharp$

Randomization	Networks	Infections	Test Statistic
$\mathbf{z}_1$	$\overline{\mathbf{A}}(\mathbf{z}_1) = \overline{\mathbf{A}}$	$\overline{\mathbf{Y}}(\mathbf{z}_1) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_1, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
$\mathbf{z}_2$	$\overline{\mathbf{A}}(\mathbf{z}_2) = \overline{\mathbf{A}}$	$\overline{\mathbf{Y}}(\mathbf{z}_2) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_2, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
...	...	...	...
$\mathbf{z}_{\lvert\mathcal{Z}\rvert}$	$\overline{\mathbf{A}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = \overline{\mathbf{A}}$	$\overline{\mathbf{Y}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_{\lvert\mathcal{Z}\rvert}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$

Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?" We can answer this using a **randomization test** (Fisher, 1936; Alexandria et al., 2023; Zhang and Zhao, 2023; Ritzwoller et al., 2024)

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under $H_0^\sharp$
  - This is possible since $\overline{\mathbf{A}}(\mathbf{z}), \overline{\mathbf{Y}}(\mathbf{z})$ are **imputable** under $H_0^\sharp$

Randomization	Networks	Infections	Test Statistic
$\mathbf{z}_1$	$\overline{\mathbf{A}}(\mathbf{z}_1) = \overline{\mathbf{A}}$	$\overline{\mathbf{Y}}(\mathbf{z}_1) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_1, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
$\mathbf{z}_2$	$\overline{\mathbf{A}}(\mathbf{z}_2) = \overline{\mathbf{A}}$	$\overline{\mathbf{Y}}(\mathbf{z}_2) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_2, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
$\dots$	$\dots$	$\dots$	$\dots$
$\mathbf{z}_{\lvert\mathcal{Z}\rvert}$	$\overline{\mathbf{A}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = \overline{\mathbf{A}}$	$\overline{\mathbf{Y}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_{\lvert\mathcal{Z}\rvert}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$

Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?" We can answer this using a **randomization test** (Fisher, 1936; Alexandria et al., 2023; Zhang and Zhao, 2023; Ritzwoller et al., 2024)

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under $H_0^\sharp$
  - This is possible since $\overline{\mathbf{A}}(\mathbf{z}), \overline{\mathbf{Y}}(\mathbf{z})$ are **imputable** under $H_0^\sharp \implies H_0^\sharp$ is **sharp**

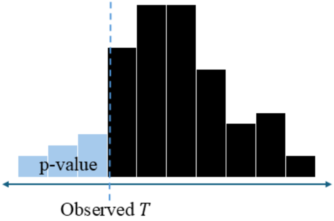
Randomization	Networks	Infections	Test Statistic
$\mathbf{z}_1$	$\overline{\mathbf{A}}(\mathbf{z}_1) = \overline{\mathbf{A}}$	$\overline{\mathbf{Y}}(\mathbf{z}_1) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_1, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
$\mathbf{z}_2$	$\overline{\mathbf{A}}(\mathbf{z}_2) = \overline{\mathbf{A}}$	$\overline{\mathbf{Y}}(\mathbf{z}_2) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_2, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
...	...	...	...
$\mathbf{z}_{\lvert\mathcal{Z}\rvert}$	$\overline{\mathbf{A}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = \overline{\mathbf{A}}$	$\overline{\mathbf{Y}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_{\lvert\mathcal{Z}\rvert}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$

Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?" We can answer this using a **randomization test** (Fisher, 1936; Alexandria et al., 2023; Zhang and Zhao, 2023; Ritzwoller et al., 2024)

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under $H_0^\sharp$
  - This is possible since $\overline{\mathbf{A}}(\mathbf{z}), \overline{\mathbf{Y}}(\mathbf{z})$ are **imputable** under $H_0^\sharp \implies H_0^\sharp$ is **sharp**
- Compute a p-value: $\rho^\sharp = F_T(T|H_0^\sharp)$, where

$$F_T(t|H_0^\sharp) = \sum_{\mathbf{z} \in \mathcal{Z}} r(\mathbf{z}) \mathbb{1}\{T(\mathbf{z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}}) \leq t\}$$



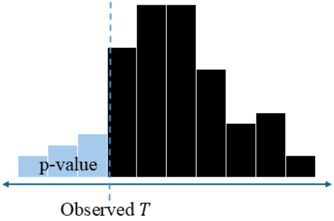
Testing $H_0^\sharp$

"How unlikely are the observed data under $H_0^\sharp$?" We can answer this using a **randomization test** (Fisher, 1936; Alexandria et al., 2023; Zhang and Zhao, 2023; Ritzwoller et al., 2024)

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under $H_0^\sharp$
  - This is possible since $\overline{\mathbf{A}}(\mathbf{z}), \overline{\mathbf{Y}}(\mathbf{z})$ are **imputable** under $H_0^\sharp \implies H_0^\sharp$ is **sharp**
- Compute a p-value: $\rho^\sharp = F_T(T|H_0^\sharp)$, where

$$F_T(t|H_0^\sharp) = \sum_{\mathbf{z} \in \mathcal{Z}} r(\mathbf{z}) \mathbb{1}\{T(\mathbf{z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}}) \leq t\}$$

- Reject $H_0^\sharp$ if $\rho^\sharp \leq 0.05$



Testing $H_0^\sharp$

- This test of $H_0^\sharp$ controls the **type I error rate** exactly:

$$\Pr\left(\rho^\sharp \leq \alpha | H_0^\sharp\right) \leq \alpha$$

for any $\alpha \in [0, 1]$, for any sample size, and for any choice of test statistic

Testing $H_0^\sharp$

- This test of $H_0^\sharp$ controls the **type I error rate** exactly:

$$\Pr\left(\rho^\sharp \leq \alpha | H_0^\sharp\right) \leq \alpha$$

  for any $\alpha \in [0, 1]$, for any sample size, and for any choice of test statistic

- The **power** of this test:

$$\Pr\left(\rho^\sharp \leq \alpha | H_1\right)$$

  depends on the choice of test statistic

Testing $H_0^\sharp$

- This test of $H_0^\sharp$ controls the **type I error rate** exactly:

$$\Pr\left(\rho^\sharp \leq \alpha | H_0^\sharp\right) \leq \alpha$$

  for any $\alpha \in [0, 1]$, for any sample size, and for any choice of test statistic

- The **power** of this test:

$$\Pr\left(\rho^\sharp \leq \alpha | H_1\right)$$

  depends on the choice of test statistic
  - In general, T should be chosen such that small values indicate departures from $H_0^\sharp$

Testing H_0^Y

"How unlikely are the observed data under H_0^Y?"

Testing H_0^Y

"How unlikely are the observed data under H_0^Y?"

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$

Testing H_0^Y

"How unlikely are the observed data under H_0^Y?"

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under H_0^Y

Testing H_0^Y

"How unlikely are the observed data under H_0^Y?"

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under H_0^Y

Randomization	Networks	Infections	Test Statistic
$\mathbf{z}_1$			
$\mathbf{z}_2$			
$\cdots$			
$\mathbf{z}_{\|\mathcal{Z}\|}$			

Testing H_0^Y

"How unlikely are the observed data under H_0^Y?"

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under H_0^Y

Randomization	Networks	Infections	Test Statistic
$\mathbf{z}_1$	$\overline{\mathbf{A}}(\mathbf{z}_1) =?$	$\overline{\mathbf{Y}}(\mathbf{z}_1) =?$	$T(\mathbf{z}_1, ?, ?)$
$\mathbf{z}_2$	$\overline{\mathbf{A}}(\mathbf{z}_2) =?$	$\overline{\mathbf{Y}}(\mathbf{z}_1) =?$	$T(\mathbf{z}_1, ?, ?)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\mathbf{z}_{\lvert\mathcal{Z}\rvert}$	$\overline{\mathbf{A}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) =?$	$\overline{\mathbf{Y}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) =?$	$T(\mathbf{z}_{\lvert\mathcal{Z}\rvert}, ?, ?)$

Testing H_0^Y

"How unlikely are the observed data under H_0^Y?"

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under H_0^Y

Randomization	Networks	Infections	Test Statistic
$\mathbf{z}_1$	$\overline{\mathbf{A}}(\mathbf{z}_1) =?$	$\overline{\mathbf{Y}}(\mathbf{z}_1) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_1, ?, \overline{\mathbf{Y}})$
$\mathbf{z}_2$	$\overline{\mathbf{A}}(\mathbf{z}_2) =?$	$\overline{\mathbf{Y}}(\mathbf{z}_2) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_2, ?, \overline{\mathbf{Y}})$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\mathbf{z}_{\lvert\mathcal{Z}\rvert}$	$\overline{\mathbf{A}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) =?$	$\overline{\mathbf{Y}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_{\lvert\mathcal{Z}\rvert}, ?, \overline{\mathbf{Y}})$

Testing H_0^Y

"How unlikely are the observed data under H_0^Y?"

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under H_0^Y
  - This is not possible since $\overline{\mathbf{A}}(\mathbf{z})$ are **not imputable** under H_0^Y

Randomization	Networks	Infections	Test Statistic
$\mathbf{z}_1$	$\overline{\mathbf{A}}(\mathbf{z}_1) = ?$	$\overline{\mathbf{Y}}(\mathbf{z}_1) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_1, ?, \overline{\mathbf{Y}})$
$\mathbf{z}_2$	$\overline{\mathbf{A}}(\mathbf{z}_2) = ?$	$\overline{\mathbf{Y}}(\mathbf{z}_2) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_2, ?, \overline{\mathbf{Y}})$
...	...	...	...
$\mathbf{z}_{\lvert\mathcal{Z}\rvert}$	$\overline{\mathbf{A}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = ?$	$\overline{\mathbf{Y}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_{\lvert\mathcal{Z}\rvert}, ?, \overline{\mathbf{Y}})$

Testing H_0^Y

"How unlikely are the observed data under H_0^Y?"

- Choose a test statistic $T = T(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}})$
- Find the distribution of T under H_0^Y
  - This is not possible since $\overline{\mathbf{A}}(\mathbf{z})$ are **not imputable** under $H_0^Y \implies H_0^Y$ is **nonsharp**

Randomization	Networks	Infections	Test Statistic
$\mathbf{z}_1$	$\overline{\mathbf{A}}(\mathbf{z}_1) = ?$	$\overline{\mathbf{Y}}(\mathbf{z}_1) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_1, ?, \overline{\mathbf{Y}})$
$\mathbf{z}_2$	$\overline{\mathbf{A}}(\mathbf{z}_2) = ?$	$\overline{\mathbf{Y}}(\mathbf{z}_2) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_2, ?, \overline{\mathbf{Y}})$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\mathbf{z}_{\lvert\mathcal{Z}\rvert}$	$\overline{\mathbf{A}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = ?$	$\overline{\mathbf{Y}}(\mathbf{z}_{\lvert\mathcal{Z}\rvert}) = \overline{\mathbf{Y}}$	$T(\mathbf{z}_{\lvert\mathcal{Z}\rvert}, ?, \overline{\mathbf{Y}})$

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp
  $\implies$ we don't know $F_T(t|H_0^Y)$ and thus cannot compute a p-value testing H_0^Y

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp
  $\implies$ we don't know $F_T(t|H_0^Y)$ and thus cannot compute a p-value testing H_0^Y

- **Idea 1**: use a **conditional randomization test** (Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022; Zhang and Zhao, 2023)

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp
  $\Longrightarrow$ we don't know $F_T(t|H_0^Y)$ and thus cannot compute a p-value testing H_0^Y

- **Idea 1**: use a **conditional randomization test** (Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022; Zhang and Zhao, 2023)
  - This involves cleverly partitioning $\mathcal{Z}$ into subsets $\mathbf{S}_1, \ldots, \mathbf{S}_M$ and restricting the randomization distribution to the subset $\mathbf{S}_m$ containing the observed $\mathbf{Z}$

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp
  $\implies$ we don't know $F_T(t|H_0^Y)$ and thus cannot compute a p-value testing H_0^Y

- **Idea 1**: use a **conditional randomization test** (Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022; Zhang and Zhao, 2023)
  - This involves cleverly partitioning $\mathcal{Z}$ into subsets $\mathbf{S}_1, \ldots, \mathbf{S}_M$ and restricting the randomization distribution to the subset $\mathbf{S}_m$ containing the observed $\mathbf{Z}$
  - Conditional randomization tests require a **partially sharp null**, but H_0^Y is **fully nonsharp** in $\overline{\mathbf{A}}(\mathbf{z})$

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp
  $\implies$ we don't know $F_T(t|H_0^Y)$ and thus cannot compute a p-value testing H_0^Y

- **Idea 1**: use a **conditional randomization test** (Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022; Zhang and Zhao, 2023)
  - This involves cleverly partitioning $\mathcal{Z}$ into subsets $\mathbf{S}_1, \ldots, \mathbf{S}_M$ and restricting the randomization distribution to the subset $\mathbf{S}_m$ containing the observed $\mathbf{Z}$
  - Conditional randomization tests require a **partially sharp null**, but H_0^Y is **fully nonsharp** in $\overline{\mathbf{A}}(\mathbf{z})$

- **Idea 2**: assume the potential networks are **stochastic**

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp
  $\implies$ we don't know $F_T(t|H_0^Y)$ and thus cannot compute a p-value testing H_0^Y

- **Idea 1**: use a **conditional randomization test** (Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022; Zhang and Zhao, 2023)
  - This involves cleverly partitioning $\mathcal{Z}$ into subsets $\mathbf{S}_1, \ldots, \mathbf{S}_M$ and restricting the randomization distribution to the subset $\mathbf{S}_m$ containing the observed $\mathbf{Z}$
  - Conditional randomization tests require a **partially sharp null**, but H_0^Y is **fully nonsharp** in $\overline{\mathbf{A}}(\mathbf{z})$

- **Idea 2**: assume the potential networks are **stochastic**
  - Define $q(\bar{\mathbf{a}}, \mathbf{z}) = \Pr\{\overline{\mathbf{A}}(\mathbf{z}) = \bar{\mathbf{a}}\}$, the PMF of $\overline{\mathbf{A}}(\mathbf{z})$

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp
  $\implies$ we don't know $F_T(t|H_0^Y)$ and thus cannot compute a p-value testing H_0^Y

- **Idea 1**: use a **conditional randomization test** (Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022; Zhang and Zhao, 2023)
  - This involves cleverly partitioning $\mathcal{Z}$ into subsets $\mathbf{S}_1, \ldots, \mathbf{S}_M$ and restricting the randomization distribution to the subset $\mathbf{S}_m$ containing the observed $\mathbf{Z}$
  - Conditional randomization tests require a **partially sharp null**, but H_0^Y is **fully nonsharp** in $\overline{\mathbf{A}}(\mathbf{z})$

- **Idea 2**: assume the potential networks are **stochastic**
  - Define $q(\bar{\mathbf{a}}, \mathbf{z}) = \Pr\{\overline{\mathbf{A}}(\mathbf{z}) = \bar{\mathbf{a}}\}$, the PMF of $\overline{\mathbf{A}}(\mathbf{z})$
  - Then

$$F_T(t|H_0^Y) = \sum_{\mathbf{z} \in \mathcal{Z}} \sum_{\bar{\mathbf{a}} \in \mathcal{A}} r(\mathbf{z}) q(\bar{\mathbf{a}}, \mathbf{z}) \mathbb{1}\{T(\mathbf{z}, \bar{\mathbf{a}}, \overline{\mathbf{Y}}) \leq t\}$$

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp
  $\implies$ we don't know $F_T(t|H_0^Y)$ and thus cannot compute a p-value testing H_0^Y

- **Idea 1**: use a **conditional randomization test** (Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022; Zhang and Zhao, 2023)
  - This involves cleverly partitioning $\mathcal{Z}$ into subsets $\mathbf{S}_1, \ldots, \mathbf{S}_M$ and restricting the randomization distribution to the subset $\mathbf{S}_m$ containing the observed $\mathbf{Z}$
  - Conditional randomization tests require a **partially sharp null**, but H_0^Y is **fully nonsharp** in $\overline{\mathbf{A}}(\mathbf{z})$

- **Idea 2**: assume the potential networks are **stochastic**
  - Define $q(\bar{\mathbf{a}}, \mathbf{z}) = \Pr\{\overline{\mathbf{A}}(\mathbf{z}) = \bar{\mathbf{a}}\}$, the PMF of $\overline{\mathbf{A}}(\mathbf{z})$
  - Then

$$F_T(t|H_0^Y) = \sum_{\mathbf{z} \in \mathcal{Z}} \sum_{\bar{\mathbf{a}} \in \mathcal{A}} r(\mathbf{z}) q(\bar{\mathbf{a}}, \mathbf{z}) \mathbb{1}\{T(\mathbf{z}, \bar{\mathbf{a}}, \overline{\mathbf{Y}}) \leq t\}$$

  - $\rho^Y = F_T(T|H_0^Y)$

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp
  $\implies$ we don't know $F_T(t|H_0^Y)$ and thus cannot compute a p-value testing H_0^Y

- **Idea 1**: use a **conditional randomization test** (Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022; Zhang and Zhao, 2023)
  - This involves cleverly partitioning $\mathcal{Z}$ into subsets $\mathbf{S}_1, \ldots, \mathbf{S}_M$ and restricting the randomization distribution to the subset $\mathbf{S}_m$ containing the observed $\mathbf{Z}$
  - Conditional randomization tests require a **partially sharp null**, but H_0^Y is **fully nonsharp** in $\overline{\mathbf{A}}(\mathbf{z})$

- **Idea 2**: assume the potential networks are **stochastic**
  - Define $q(\overline{\mathbf{a}}, \mathbf{z}) = \Pr\{\overline{\mathbf{A}}(\mathbf{z}) = \overline{\mathbf{a}}\}$, the PMF of $\overline{\mathbf{A}}(\mathbf{z})$
  - Then

$$F_T(t|H_0^Y) = \sum_{\mathbf{z}\in\mathcal{Z}} \sum_{\overline{\mathbf{a}}\in\mathcal{A}} r(\mathbf{z}) q(\overline{\mathbf{a}}, \mathbf{z}) \mathbb{1}\{T(\mathbf{z}, \overline{\mathbf{a}}, \overline{\mathbf{Y}}) \leq t\}$$

  - $\rho^Y = F_T(T|H_0^Y)$
  - If/when q is unknown, estimate it

Testing H_0^Y

- **Problem**: H_0^Y is nonsharp
  $\Longrightarrow$ we don't know $F_T(t|H_0^Y)$ and thus cannot compute a p-value testing H_0^Y

- **Idea 1**: use a **conditional randomization test** (Athey et al., 2018; Basse et al., 2019; Puelz et al., 2022; Zhang and Zhao, 2023)
  - This involves cleverly partitioning $\mathcal{Z}$ into subsets $\mathbf{S}_1, \ldots, \mathbf{S}_M$ and restricting the randomization distribution to the subset $\mathbf{S}_m$ containing the observed $\mathbf{Z}$
  - Conditional randomization tests require a **partially sharp null**, but H_0^Y is **fully nonsharp** in $\overline{\mathbf{A}}(\mathbf{z})$

- **Idea 2**: assume the potential networks are **stochastic**
  - Define $q(\overline{\mathbf{a}}, \mathbf{z}) = \Pr\{\overline{\mathbf{A}}(\mathbf{z}) = \overline{\mathbf{a}}\}$, the PMF of $\overline{\mathbf{A}}(\mathbf{z})$
  - Then

$$F_T(t|H_0^Y) = \sum_{\mathbf{z}\in\mathcal{Z}} \sum_{\overline{\mathbf{a}}\in\mathcal{A}} r(\mathbf{z}) q(\overline{\mathbf{a}}, \mathbf{z}) \mathbb{1}\{T(\mathbf{z}, \overline{\mathbf{a}}, \overline{\mathbf{Y}}) \leq t\}$$

  - $\rho^Y = F_T(T|H_0^Y)$
  - If/when q is unknown, estimate it
    - E.g., using a **Separable Temporal Exponential Random Graph Model**
    - Use estimated q to obtain $\widehat{\rho}^Y$

Testing H_0^Y

Testing Procedure:

1. Estimate the unknown parameter θ in the PMF $q(\overline{\mathbf{a}}, \mathbf{z}; \theta)$ of $\overline{\mathbf{A}}(\mathbf{z})$

Testing H_0^Y

Testing Procedure:

1. Estimate the unknown parameter θ in the PMF $q(\overline{\mathbf{a}}, \mathbf{z}; \theta)$ of $\overline{\mathbf{A}}(\mathbf{z})$
2. Approximate $F_T(t|H_0^Y)$ using B Monte-Carlo replicates:
   for $b = 1$ to B **do:**

Testing H_0^Y

Testing Procedure:

1. Estimate the unknown parameter θ in the PMF $q(\bar{\mathbf{a}}, \mathbf{z}; \theta)$ of $\overline{\mathbf{A}}(\mathbf{z})$
2. Approximate $F_T(t|H_0^Y)$ using B Monte-Carlo replicates:
   for $b = 1$ to B **do:**
   1. generate randomization vector $\mathbf{z}_b$

Testing H_0^Y

Testing Procedure:

1. Estimate the unknown parameter θ in the PMF $q(\overline{\mathbf{a}}, \mathbf{z}; \theta)$ of $\overline{\mathbf{A}}(\mathbf{z})$
2. Approximate $F_T(t|H_0^Y)$ using B Monte-Carlo replicates:
   for $b = 1$ to B **do:**
   1. generate randomization vector $\mathbf{z}_b$
   2. generate networks $\overline{\mathbf{a}}_b$ from $q(\overline{\mathbf{a}}, \mathbf{z}_b; \widehat{\theta})$

Testing H_0^Y

Testing Procedure:

1. Estimate the unknown parameter θ in the PMF $q(\bar{\mathbf{a}}, \mathbf{z}; \theta)$ of $\overline{\mathbf{A}}(\mathbf{z})$
2. Approximate $F_T(t|H_0^Y)$ using B Monte-Carlo replicates:
   for $b = 1$ to B **do:**
   1. generate randomization vector $\mathbf{z}_b$
   2. generate networks $\bar{\mathbf{a}}_b$ from $q(\bar{\mathbf{a}}, \mathbf{z}_b; \widehat{\theta})$
   3. compute the test statistic $T_b = T(\mathbf{z}_b, \bar{\mathbf{a}}_b, \mathbf{Y})$

Testing H_0^Y

Testing Procedure:

1. Estimate the unknown parameter θ in the PMF $q(\bar{\mathbf{a}}, \mathbf{z}; \theta)$ of $\overline{\mathbf{A}}(\mathbf{z})$
2. Approximate $F_T(t|H_0^Y)$ using B Monte-Carlo replicates:
   for $b = 1$ to B **do:**
   1. generate randomization vector $\mathbf{z}_b$
   2. generate networks $\bar{\mathbf{a}}_b$ from $q(\bar{\mathbf{a}}, \mathbf{z}_b; \widehat{\theta})$
   3. compute the test statistic $T_b = T(\mathbf{z}_b, \bar{\mathbf{a}}_b, \mathbf{Y})$
3. compute the plug-in p-value testing H_0^Y:

$$\widehat{\rho}_B^Y = B^{-1} \sum_{b=1}^{B} \mathbb{1}(T_b \leq T)$$

Testing H_0^Y

Testing Procedure:

1. Estimate the unknown parameter θ in the PMF $q(\bar{\mathbf{a}}, \mathbf{z}; \theta)$ of $\overline{\mathbf{A}}(\mathbf{z})$
2. Approximate $F_T(t|H_0^Y)$ using B Monte-Carlo replicates:
   for $b = 1$ to B **do:**
   1. generate randomization vector $\mathbf{z}_b$
   2. generate networks $\bar{\mathbf{a}}_b$ from $q(\bar{\mathbf{a}}, \mathbf{z}_b; \widehat{\theta})$
   3. compute the test statistic $T_b = T(\mathbf{z}_b, \bar{\mathbf{a}}_b, \mathbf{Y})$
3. compute the plug-in p-value testing H_0^Y:

$$\widehat{\rho}_B^Y = B^{-1} \sum_{b=1}^{B} \mathbb{1}(T_b \leq T)$$

Proposition: This procedure will **asymptotically** control the type I error rate

Simulation Study

- $n \in \{24, 48\}$ students
- $\tau \in \{5, 10\}$ weeks
- 50:50 cluster randomization with 12 equal-sized clusters

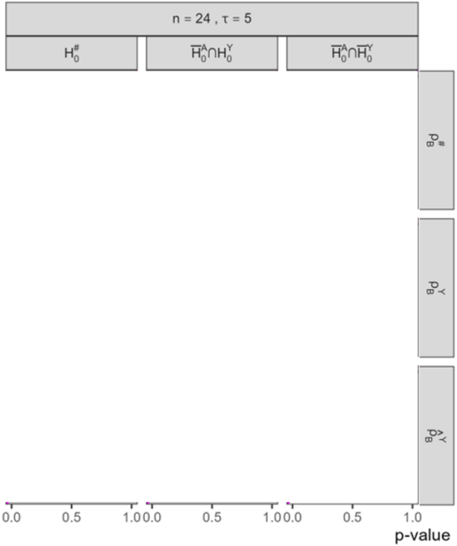
Simulation Study

- $n \in \{24, 48\}$ students
- $\tau \in \{5, 10\}$ weeks
- 50:50 cluster randomization with 12 equal-sized clusters
- Three scenarios:
    1. $H_0^\sharp$: no effect of intervention on networks or infection
    2. $\overline{H}_0^A \cap H_0^Y$: no effect of intervention on infection
    3. $\overline{H}_0^A \cap \overline{H}_0^Y$: effect of intervention on both networks and infection

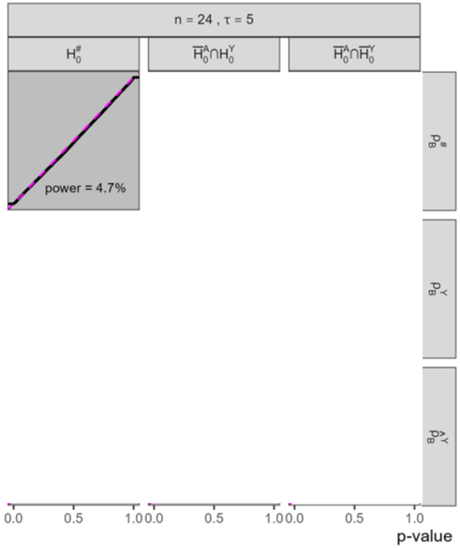
Simulation Study

- $n \in \{24, 48\}$ students
- $\tau \in \{5, 10\}$ weeks
- 50:50 cluster randomization with 12 equal-sized clusters
- Three scenarios:
  1. $H_0^\sharp$: no effect of intervention on networks or infection
  2. $\overline{H}_0^A \cap H_0^Y$: no effect of intervention on infection
  3. $\overline{H}_0^A \cap \overline{H}_0^Y$: effect of intervention on both networks and infection
- Three testing procedures:
  1. $\rho_B^\sharp$: testing $H_0^\sharp$
  2. ρ_B^Y: testing H_0^Y using known q
  3. $\widehat{\rho}_B^Y$: testing H_0^Y using estimated q

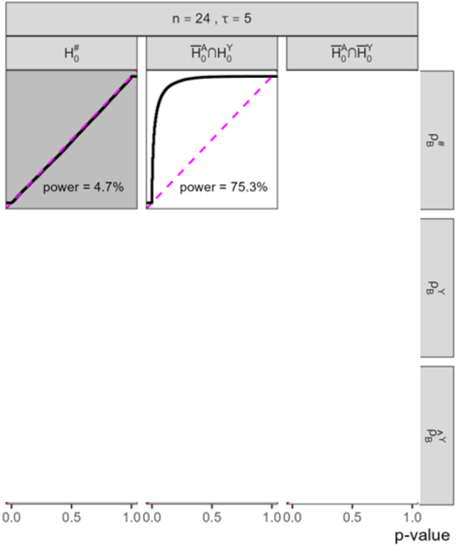
Simulation Results



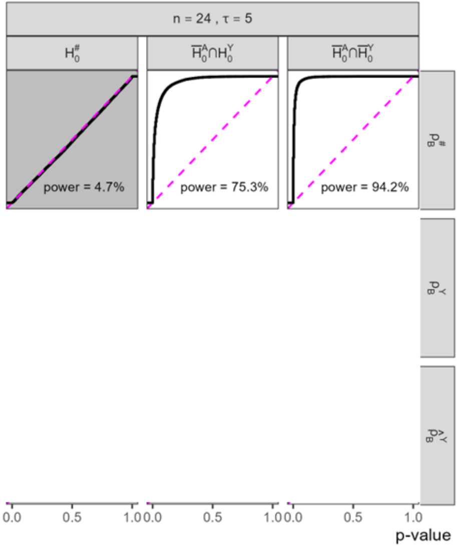
Simulation Results



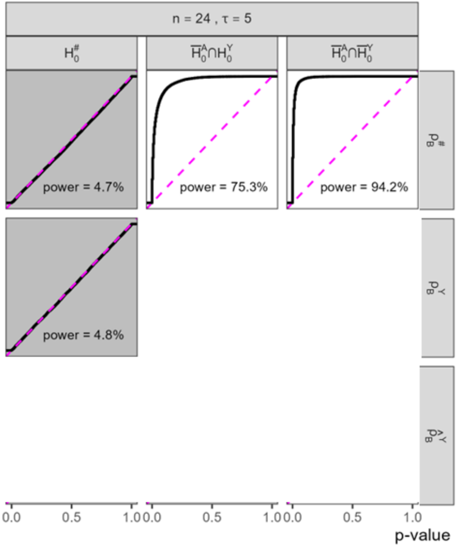
Simulation Results



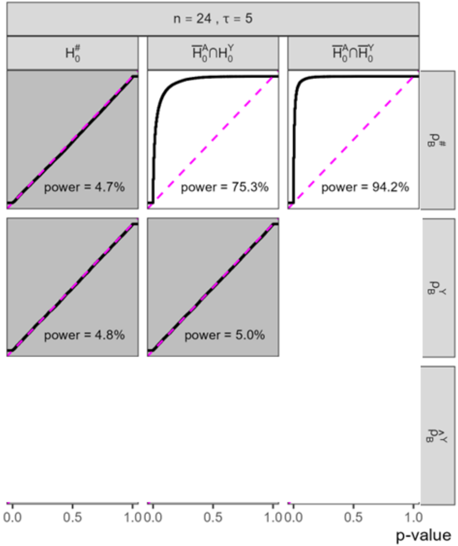
Simulation Results



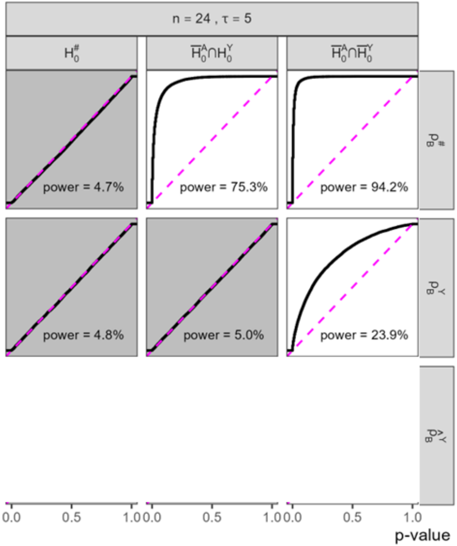
Simulation Results



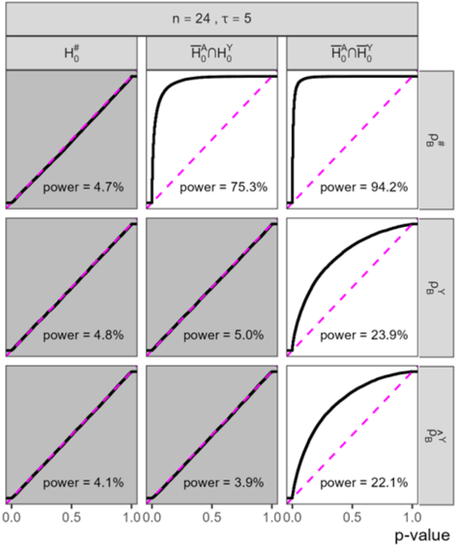
Simulation Results



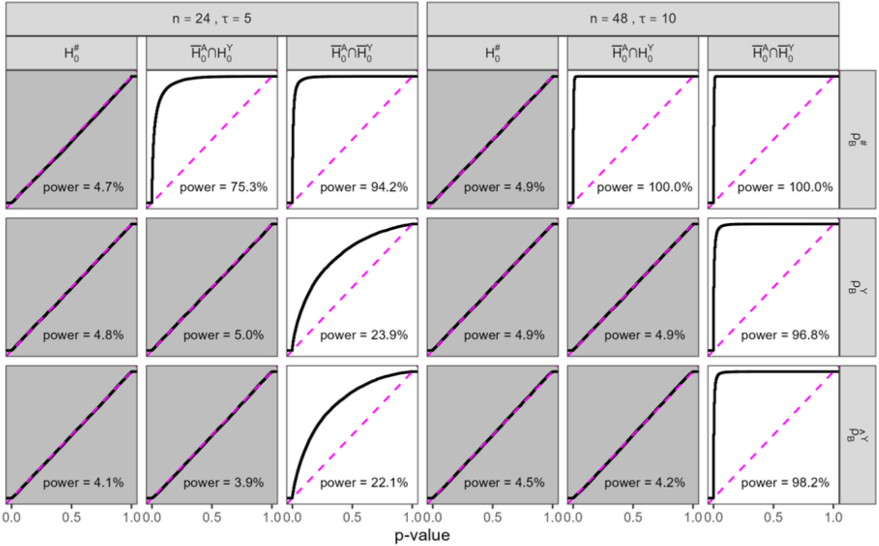
Simulation Results



Simulation Results



Simulation Results

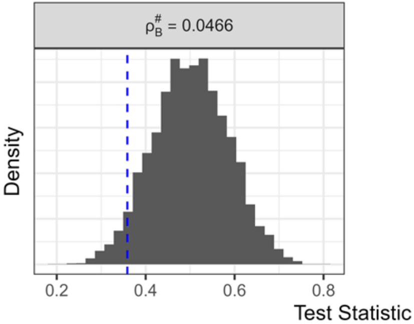


eX-FLU: Hypothesis Test Results

- Test statistic (proportion of possible transmission events attributable to students in the intervention group): $T = 0.359$.

eX-FLU: Hypothesis Test Results

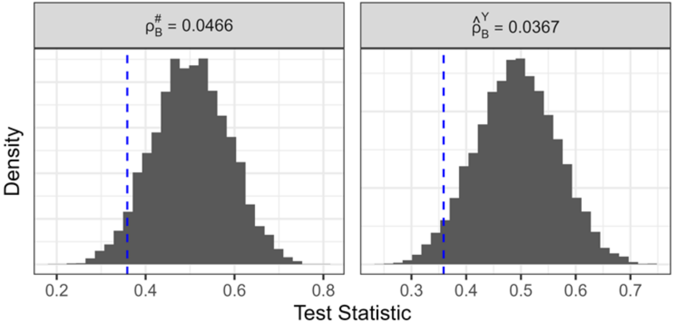
- Test statistic (proportion of possible transmission events attributable to students in the intervention group): $T = 0.359$.



- Encouragement to isolate affects the social network and/or influenze-like illness ($\rho_B^{\sharp} = 0.0466$)

eX-FLU: Hypothesis Test Results

- Test statistic (proportion of possible transmission events attributable to students in the intervention group): $T = 0.359$.



- Encouragement to isolate affects the social network and/or influenze-like illness ($\rho_B^\sharp = 0.0466$)
- Encouragement to isolate specifically affects influenze-like illness ($\widehat{\rho}_B^Y = 0.0367$)

Summary of Project 2

- Developed a new randomization-based inference procedure
  - Established theoretical properties (asymptotic control of Type I error)
  - Demonstrated empirical performance through simulations
- Applied the method to the **eX-FLU trial** to find a protective effect of an encouragement-to-isolate intervention on a college campus
- **Status**: preparing submission to *PNAS* (target: December)

Project 3 : Causal Inference from Cluster Randomized Trials with Differential Nonresponse

Brian Richardson, Bonnie Shook-Sa, Michael Hudgens

PopART Trial

- **PopART**: large **cluster randomized trial** designed to evaluate a combination **HIV prevention intervention** in Zambia and South Africa (Hayes et al., 2014)

Hayes *et al. Trials* 2014, **15**:57
http://www.trialsjournal.com/content/15/1/57

TRIALS

STUDY PROTOCOL **Open Access**

HPTN 071 (PopART): Rationale and design of a cluster-randomised trial of the population impact of an HIV combination prevention intervention including universal testing and treatment – a study protocol for a cluster randomised trial

Richard Hayes[1], Helen Ayles[2,3], Nulda Beyers[4], Kalpana Sabapathy[1*], Sian Floyd[1], Kwame Shanaube[3], Peter Bock[4], Sam Griffith[5], Ayana Moore[5], Deborah Watson-Jones[2], Christophe Fraser[6], Sten H Vermund[7], Sarah Fidler[8] and The HPTN 071 (PopART) Study Team

PopART Trial

- **PopART**: large **cluster randomized trial** designed to evaluate a combination **HIV prevention intervention** in Zambia and South Africa (Hayes et al., 2014)
- **Intervention**:
  - Arm A: annual home-based HIV testing, promotion of medical male circumcision for HIV-negative men, and offer of immediate ART for those testing HIV-positive
  - Arm B: same as Arm A, but ART initiation following national guidelines
  - Arm C: standard of care

Hayes *et al. Trials* 2014, **15**:57
http://www.trialsjournal.com/content/15/1/57

TRIALS

STUDY PROTOCOL **Open Access**

HPTN 071 (PopART): Rationale and design of a cluster-randomised trial of the population impact of an HIV combination prevention intervention including universal testing and treatment – a study protocol for a cluster randomised trial

Richard Hayes[1], Helen Ayles[2,3], Nulda Beyers[4], Kalpana Sabapathy[1*], Sian Floyd[1], Kwame Shanaube[3], Peter Bock[4], Sam Griffith[5], Ayana Moore[5], Deborah Watson-Jones[2], Christophe Fraser[6], Sten H Vermund[7], Sarah Fidler[8] and The HPTN 071 (PopART) Study Team

PopART Trial

- **PopART**: large **cluster randomized trial** designed to evaluate a combination **HIV prevention intervention** in Zambia and South Africa (Hayes et al., 2014)
- **Intervention**:
  - Arm A: annual home-based HIV testing, promotion of medical male circumcision for HIV-negative men, and offer of immediate ART for those testing HIV-positive
  - Arm B: same as Arm A, but ART initiation following national guidelines
  - Arm C: standard of care
- **Outcome**: HIV Incidence

Hayes *et al. Trials* 2014, **15**:57
http://www.trialsjournal.com/content/15/1/57

TRIALS

STUDY PROTOCOL Open Access

HPTN 071 (PopART): Rationale and design of a cluster-randomised trial of the population impact of an HIV combination prevention intervention including universal testing and treatment – a study protocol for a cluster randomised trial

Richard Hayes[1], Helen Ayles[2,3], Nulda Beyers[4], Kalpana Sabapathy[1*], Sian Floyd[1], Kwame Shanaube[3], Peter Bock[4], Sam Griffith[5], Ayana Moore[5], Deborah Watson-Jones[2], Christophe Fraser[6], Sten H Vermund[7], Sarah Fidler[8] and The HPTN 071 (PopART) Study Team

PopART Trial

- **Initial Findings**: no protective effect of the Arm A intervention (Hayes et al., 2019)
  - "unanticipated and not consistent with the data on viral suppression"

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JULY 18, 2019 VOL. 381 NO. 3

Effect of Universal Testing and Treatment on HIV Incidence — HPTN 071 (PopART)

R.J. Hayes, D. Donnell, S. Floyd, N. Mandla, J. Bwalya, K. Sabapathy, B. Yang, M. Phiri, A. Schaap, S.H. Eshleman, E. Piwowar-Manning, B. Kosloff, A. James, T. Skalland, E. Wilson, L. Emel, D. Macleod, R. Dunbar, M. Simwinga, N. Makola, V. Bond, G. Hoddinott, A. Moore, S. Griffith, N. Deshmane Sista, S.H. Vermund, W. El-Sadr, D.N. Burns, J.R. Hargreaves, K. Hauck, C. Fraser, K. Shanaube, P. Bock, N. Beyers, H. Ayles, and S. Fidler, for the HPTN 071 (PopART) Study Team

PopART Trial

- **Initial Findings**: no protective effect of the Arm A intervention (Hayes et al., 2019)
  - "unanticipated and not consistent with the data on viral suppression"
- **Later Analyses**: identified possibly differential nonresponse that could bias trial results (Shook-Sa et al., 2025+)
  - Men were less likely to respond (i.e., have data collected)
  - The intervention appeared more effective among men

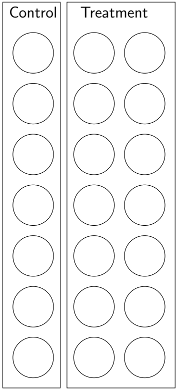
The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JULY 18, 2019 VOL. 381 NO. 3

Effect of Universal Testing and Treatment on HIV Incidence — HPTN 071 (PopART)

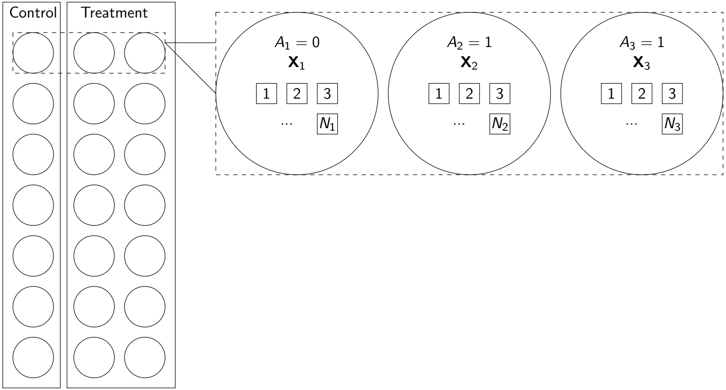
R.J. Hayes, D. Donnell, S. Floyd, N. Mandla, J. Bwalya, K. Sabapathy, B. Yang, M. Phiri, A. Schaap, S.H. Eshleman, E. Piwowar-Manning, B. Kosloff, A. James, T. Skalland, E. Wilson, L. Emel, D. Macleod, R. Dunbar, M. Simwinga, N. Makola, V. Bond, G. Hoddinott, A. Moore, S. Griffith, N. Deshmane Sista, S.H. Vermund, W. El-Sadr, D.N. Burns, J.R. Hargreaves, K. Hauck, C. Fraser, K. Shanaube, P. Bock, N. Beyers, H. Ayles, and S. Fidler, for the HPTN 071 (PopART) Study Team

PopART Trial Design



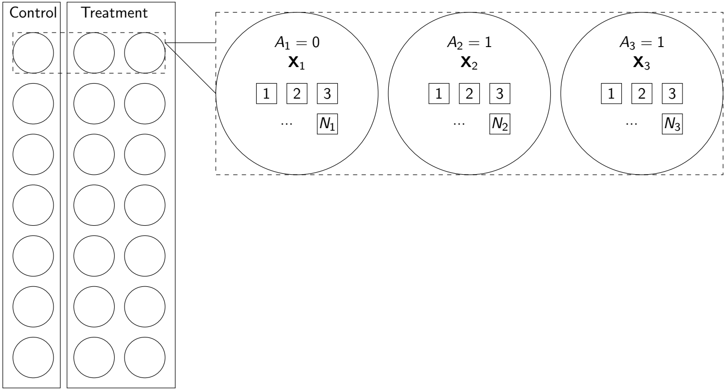
- 21 clusters randomized, each with:

PopART Trial Design



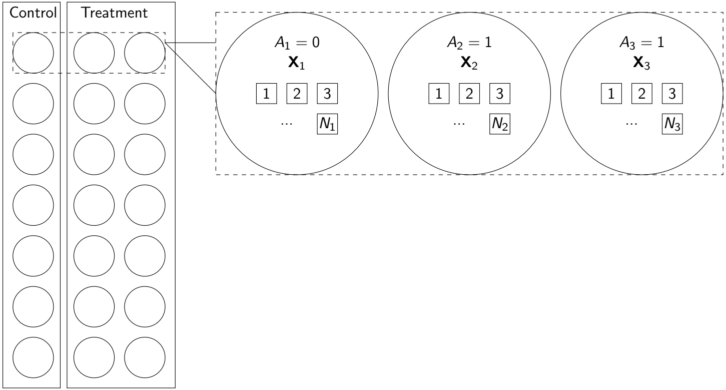
- 21 clusters randomized, each with:
  - treatment A_j

PopART Trial Design



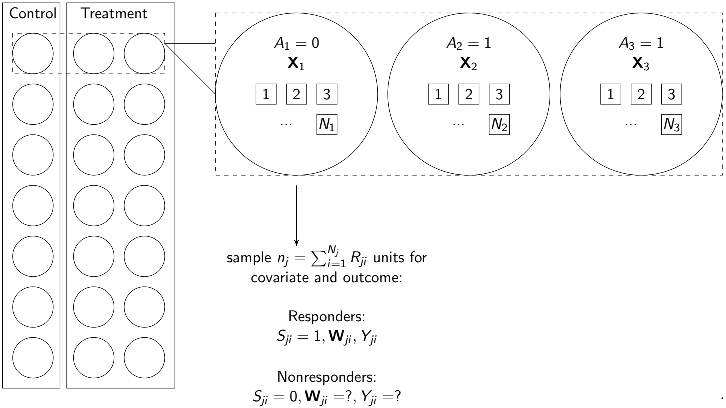
- 21 clusters randomized, each with:
  - treatment A_j
  - covariates $\mathbf{X}_j$

PopART Trial Design



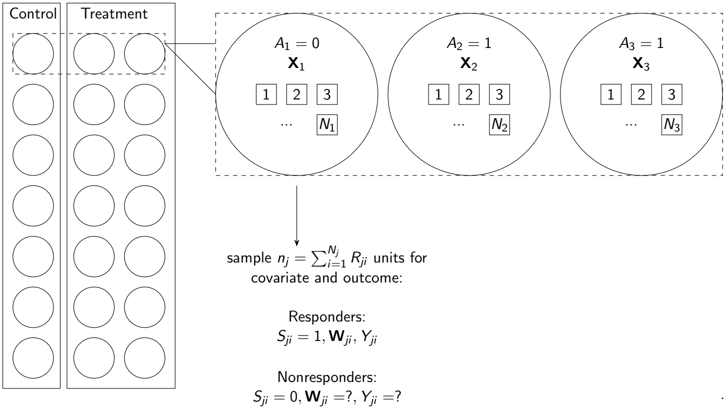
- 21 clusters randomized, each with:
  - treatment A_j
  - covariates $\mathbf{X}_j$
  - size N_j

PopART Trial Design



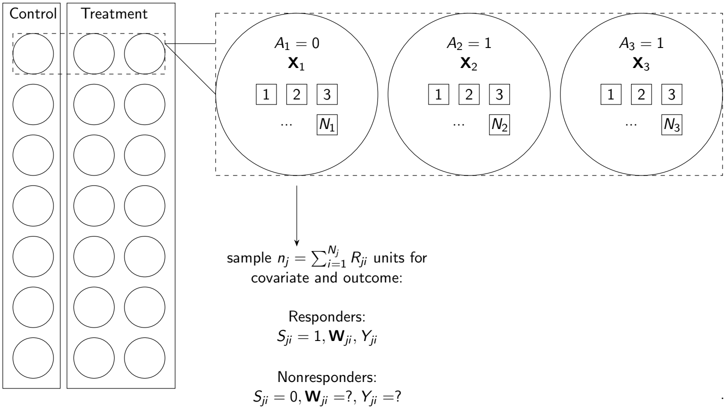
- 21 clusters randomized, each with:
  - treatment A_j
  - covariates $\mathbf{X}_j$
  - size N_j
- The trial randomly sampled n_j of N_j individuals:

PopART Trial Design



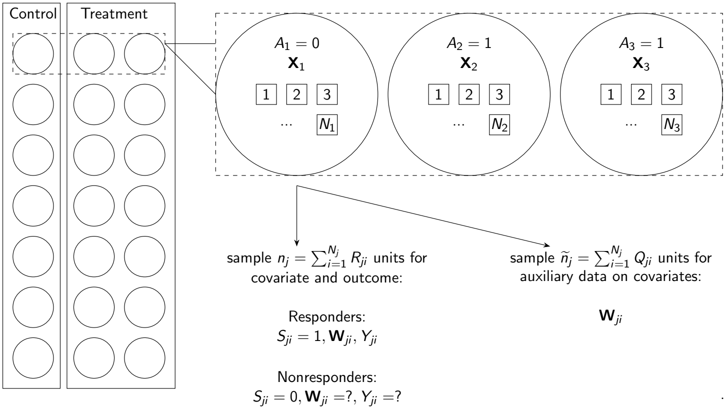
- 21 clusters randomized, each with:
  - treatment A_j
  - covariates $\mathbf{X}_j$
  - size N_j
- The trial randomly sampled n_j of N_j individuals:
  - among **responders** $(S_{ji} = 1)$:

PopART Trial Design



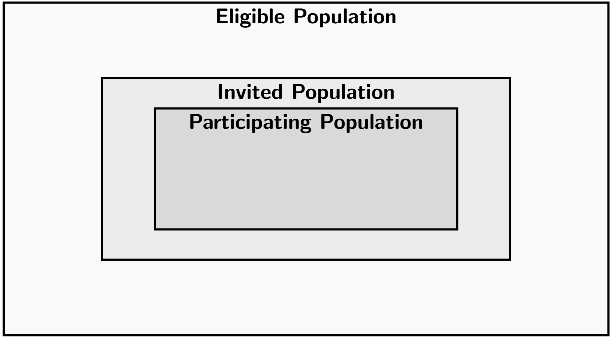
- 21 clusters randomized, each with:
  - treatment A_j
  - covariates $\mathbf{X}_j$
  - size N_j
- The trial randomly sampled n_j of N_j individuals:
  - among **responders** $(S_{ji} = 1)$:
  - covariates $\mathbf{W}_{ji}$
  - outcome Y_{ji}

PopART Trial Design



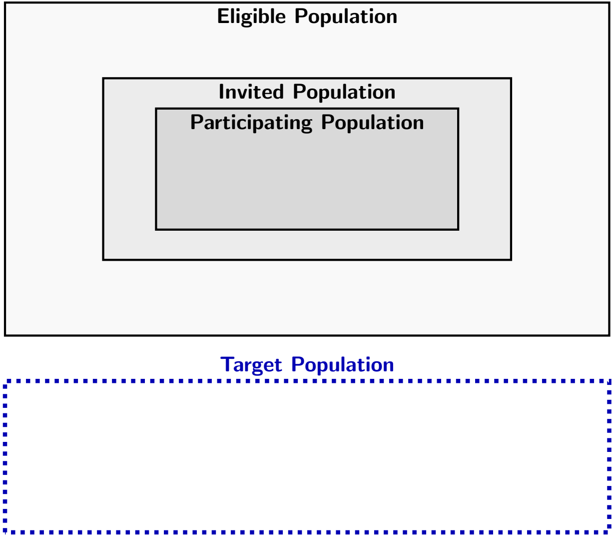
- 21 clusters randomized, each with:
  - treatment A_j
  - covariates $\mathbf{X}_j$
  - size N_j
- The trial randomly sampled n_j of N_j individuals:
  - among **responders** ($S_{ji} = 1$):
  - covariates $\mathbf{W}_{ji}$
  - outcome Y_{ji}
- Auxiliary data available for $\widetilde{n}_j$ of N_j individuals
  - covariates $\mathbf{W}_{ji}$

Extending Inference from Randomized Controlled Trials



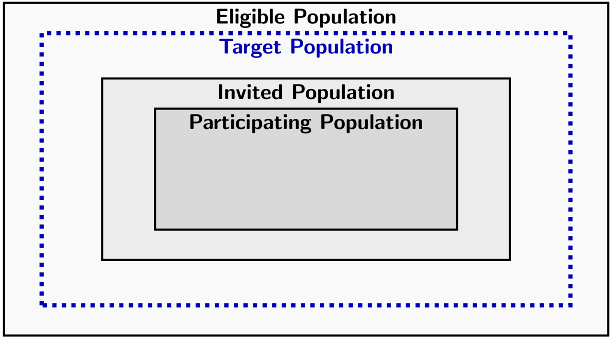
- Three populations Dahabreh and Hernán (2019):

Extending Inference from Randomized Controlled Trials



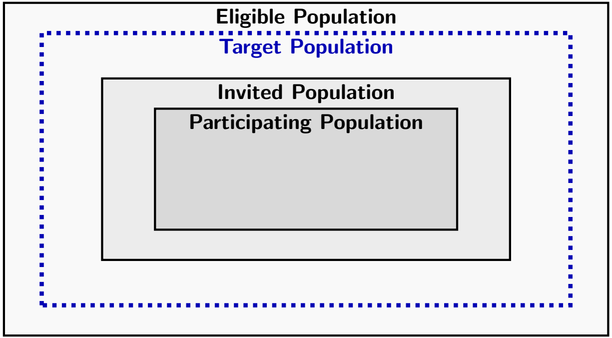
- Three populations Dahabreh and Hernán (2019):
- **Transportability**: target population and eligible population disjoint

Extending Inference from Randomized Controlled Trials



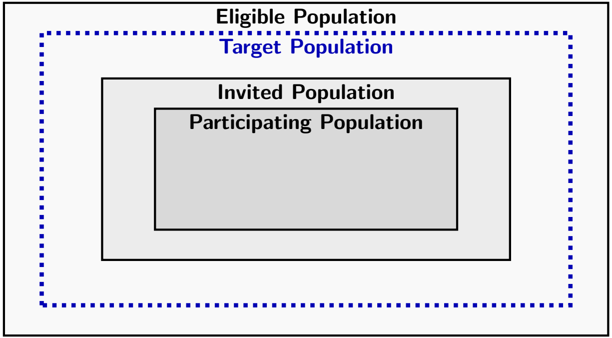
- Three populations Dahabreh and Hernán (2019):
- **Transportability**: target population and eligible population disjoint
- **Generalizability**: target population overlaps with eligible population

Extending Inference from Randomized Controlled Trials



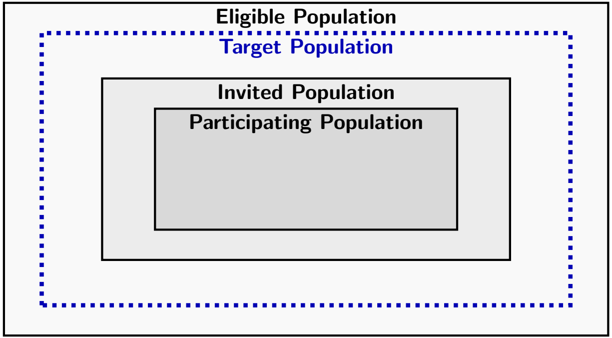
- Three populations Dahabreh and Hernán (2019):
- **Transportability**: target population and eligible population disjoint
- **Generalizability**: target population overlaps with eligible population
  - Dahabreh et al. (2019b): generalizing from participating population to eligible population using covariates from the eligible population

Extending Inference from Randomized Controlled Trials



- Three populations Dahabreh and Hernán (2019):
- **Transportability**: target population and eligible population disjoint
- **Generalizability**: target population overlaps with eligible population
  - Dahabreh et al. (2019b): generalizing from participating population to eligible population using covariates from the eligible population
  - Dahabreh et al. (2019a): generalizing from participating population to eligible population using covariates from a subset of the eligible population

Extending Inference from Randomized Controlled Trials



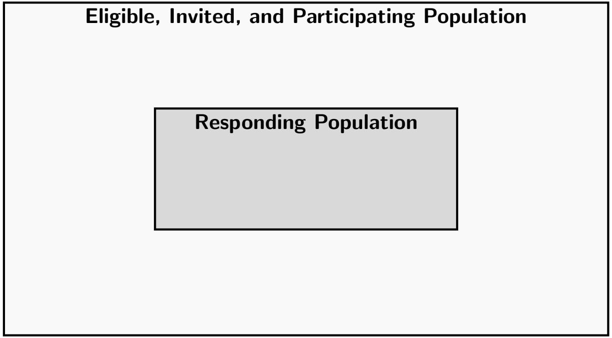
- Three populations Dahabreh and Hernán (2019):
- **Transportability**: target population and eligible population disjoint
- **Generalizability**: target population overlaps with eligible population
  - Dahabreh et al. (2019b): generalizing from participating population to eligible population using covariates from the eligible population
  - Dahabreh et al. (2019a): generalizing from participating population to eligible population using covariates from a subset of the eligible population
  - Dahabreh et al. (2020): generalizing results from participating clusters in a **cluster** RCT to eligible clusters

Participation vs Response

Eligible, Invited, and Participating Population

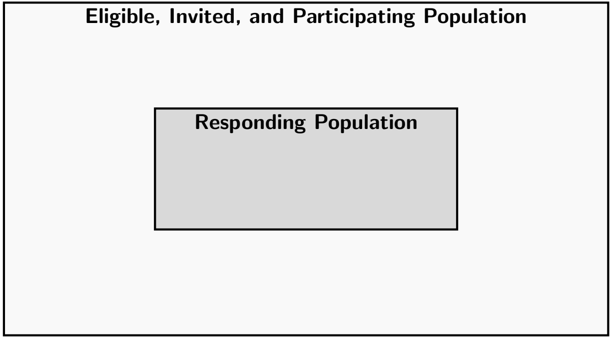
- PopART intervention was delivered at the cluster level $\implies$ all individuals **participated**

Participation vs Response



- PopART intervention was delivered at the cluster level $\Longrightarrow$ all individuals **participated**
- Not all individuals **responded**

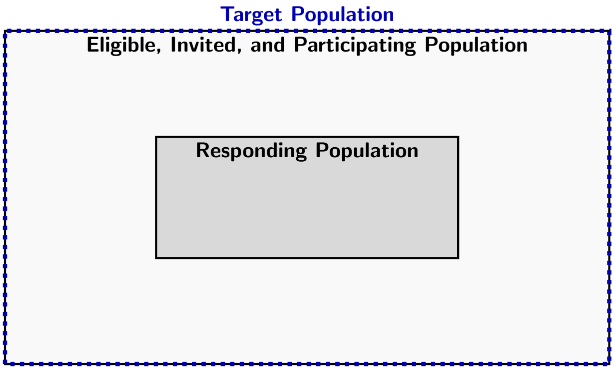
Participation vs Response



- PopART intervention was delivered at the cluster level $\implies$ all individuals **participated**
- Not all individuals **responded**
- Treatment may affect response:

$$A_j \longrightarrow S_{ji}$$

Participation vs Response



- PopART intervention was delivered at the cluster level $\implies$ all individuals **participated**
- Not all individuals **responded**
- Treatment may affect response:

$$A_j \longrightarrow S_{ji}$$

- **Goal**: generalize results to trial-eligible population using data from responders

Proposal for Project 3

- Develop estimators of the causal effect of the PopART intervention in the trial eligible population

Proposal for Project 3

- Develop estimators of the causal effect of the PopART intervention in the trial eligible population
  1. G-formula:

Proposal for Project 3

- Develop estimators of the causal effect of the PopART intervention in the trial eligible population
  1. G-formula:
     - ★ Fit an outcome regression model among responders, then marginalize over the covariate distribution in auxiliary data

Proposal for Project 3

- Develop estimators of the causal effect of the PopART intervention in the trial eligible population
  1. G-formula:
     - ★ Fit an outcome regression model among responders, then marginalize over the covariate distribution in auxiliary data
  2. Inverse probability weighting (IPW)

Proposal for Project 3

- Develop estimators of the causal effect of the PopART intervention in the trial eligible population
  1. G-formula:
     - ★ Fit an outcome regression model among responders, then marginalize over the covariate distribution in auxiliary data
  2. Inverse probability weighting (IPW)
     - ★ Weight responders to make them representative of the trial-eligible population

Proposal for Project 3

- Develop estimators of the causal effect of the PopART intervention in the trial eligible population
  1. G-formula:
     - ★ Fit an outcome regression model among responders, then marginalize over the covariate distribution in auxiliary data
  2. Inverse probability weighting (IPW)
     - ★ Weight responders to make them representative of the trial-eligible population
  3. Augmented inverse probability weighting (AIPW)

Proposal for Project 3

- Develop estimators of the causal effect of the PopART intervention in the trial eligible population
  1. G-formula:
     - ★ Fit an outcome regression model among responders, then marginalize over the covariate distribution in auxiliary data
  2. Inverse probability weighting (IPW)
     - ★ Weight responders to make them representative of the trial-eligible population
  3. Augmented inverse probability weighting (AIPW)
     - ★ Both weighting and outcome regression

Proposal for Project 3

- Develop estimators of the causal effect of the PopART intervention in the trial eligible population
  1. G-formula:
     - Fit an outcome regression model among responders, then marginalize over the covariate distribution in auxiliary data
  2. Inverse probability weighting (IPW)
     - Weight responders to make them representative of the trial-eligible population
  3. Augmented inverse probability weighting (AIPW)
     - Both weighting and outcome regression
     - Semiparametric efficient (?)

Proposal for Project 3

- Develop estimators of the causal effect of the PopART intervention in the trial eligible population
  1. G-formula:
     - ★ Fit an outcome regression model among responders, then marginalize over the covariate distribution in auxiliary data
  2. Inverse probability weighting (IPW)
     - ★ Weight responders to make them representative of the trial-eligible population
  3. Augmented inverse probability weighting (AIPW)
     - ★ Both weighting and outcome regression
     - ★ Semiparametric efficient (?)
     - ★ (Doubly) Robust (?)

Thank you! Questions?

References I

Allison E. Aiello, Amanda M. Simanek, Marisa C. Eisenberg, Alison R. Walsh, Brian Davis, Erik Volz, Caroline Cheng, Jeanette J. Rainey, Amra Uzicanin, Hongjiang Gao, Nathaniel Osgood, Dylan Knowles, Kevin Stanley, Kara Tarter, and Arnold S. Monto. Design and methods of a social network isolation study for reducing respiratory infection transmission: The eX-FLU cluster randomized trial. *Epidemics*, 15:38–55, June 2016. ISSN 17554365. doi: 10.1016/j.epidem.2016.01.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1755436516000025.

Shaina J. Alexandria, Michael G. Hudgens, and Allison E. Aiello. Assessing Intervention Effects in a Randomized Trial Within a Social Network. *Biometrics*, 79(2):1409–1419, June 2023. ISSN 0006-341X, 1541-0420. doi: 10.1111/biom.13606. URL https://academic.oup.com/biometrics/article/79/2/1409-1419/7513977.

Susan Athey, Dean Eckles, and Guido W. Imbens. Exact p -Values for Network Interference. *Journal of the American Statistical Association*, 113(521):230–240, January 2018. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2016.1241178. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2016.1241178.

G W Basse, A Feller, and P Toulis. Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494, June 2019. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asy072. URL https://academic.oup.com/biomet/article/106/2/487/5306899.

I. Dahabreh, S. Robertson, J. Steingrimsson, S. Gravenstein, and N. Joyce. Extending inferences from a cluster-randomized trial to a target population. *Health Services Research*, 55(S1):109–110, August 2020. ISSN 0017-9124, 1475-6773. doi: 10.1111/1475-6773.13486.

Issa J. Dahabreh and Miguel A. Hernán. Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology*, 34(8):719–722, August 2019. ISSN 0393-2990, 1573-7284. doi: 10.1007/s10654-019-00533-2.

References II

Issa J. Dahabreh, Miguel A. Hernan, Sarah E. Robertson, Ashley Buchanan, and Jon A. Steingrimsson. Generalizing trial findings using nested trial designs with sub-sampling of non-randomized individuals. 2019a. doi: 10.48550/ARXIV.1902.06080. URL https://arxiv.org/abs/1902.06080.

Issa J. Dahabreh, Sarah E. Robertson, Eric J. Tchetgen, Elizabeth A. Stuart, and Miguel A. Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019b. ISSN 0006-341X, 1541-0420. doi: 10.1111/biom.13009.

R. A. Fisher. The design of experiments. *Nature*, 137(3459):252–254, February 1936. ISSN 0028-0836, 1476-4687. doi: 10.1038/137252a0.

Youyi Fong, Xiaoying Shen, Vicki C Ashley, Aaron Deal, Kelly E Seaton, Chenchen Yu, Shannon P Grant, Guido Ferrari, Allan C deCamp, Robert T Bailer, et al. Modification of the association between T-cell immune responses and human immunodeficiency virus type 1 infection risk by vaccine-induced antibody responses in the HVTN 505 trial. *The Journal of Infectious Diseases*, 217(8):1280–1288, 2018.

Scott M Hammer, Magdalena E Sobieszczyk, Holly Janes, Shelly T Karuna, Mark J Mulligan, Doug Grove, Beryl A Koblin, Susan P Buchbinder, Michael C Keefer, Georgia D Tomaras, et al. Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. *New England Journal of Medicine*, 369(22):2083–2092, 2013.

Richard Hayes, Helen Ayles, Nulda Beyers, Kalpana Sabapathy, Sian Floyd, Kwame Shanaube, Peter Bock, Sam Griffith, Ayana Moore, Deborah Watson-Jones, Christophe Fraser, Sten H Vermund, Sarah Fidler, and The HPTN 071 (PopART) Study Team. Hptn 071 (popart): Rationale and design of a cluster-randomised trial of the population impact of an hiv combination prevention intervention including universal testing and treatment – a study protocol for a cluster randomised trial. *Trials*, 15(1):57, December 2014. ISSN 1745-6215. doi: 10.1186/1745-6215-15-57.

References III

Richard J. Hayes, Deborah Donnell, Sian Floyd, Nomtha Mandla, Justin Bwalya, Kalpana Sabapathy, Blia Yang, Mwelwa Phiri, Ab Schaap, Susan H. Eshleman, Estelle Piwowar-Manning, Barry Kosloff, Anelet James, Timothy Skalland, Ethan Wilson, Lynda Emel, David Macleod, Rory Dunbar, Musonda Simwinga, Nozizwe Makola, Virginia Bond, Graeme Hoddinott, Ayana Moore, Sam Griffith, Nirupama Deshmane Sista, Sten H. Vermund, Wafaa El-Sadr, David N. Burns, James R. Hargreaves, Katharina Hauck, Christophe Fraser, Kwame Shanaube, Peter Bock, Nulda Beyers, Helen Ayles, and Sarah Fidler. Effect of universal testing and treatment on hiv incidence — hptn 071 (popart). *New England Journal of Medicine*, 381(3):207–218, 2019. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa1814556.

Miguel A. Hernan and James M. Robins. *Causal inference: what if*. Taylor and Francis, Boca Raton, first edition edition, 2024. ISBN 978-1-315-37493-2.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, December 1986. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1986.10478354.

Michael G Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214508000000292.

Holly E Janes, Kristen W Cohen, Nicole Frahm, Stephen C De Rosa, Brittany Sanchez, John Hural, Craig A Magaret, Shelly Karuna, Carter Bentley, Raphael Gottardo, et al. Higher T-cell responses induced by DNA/rAd5 HIV-1 preventive vaccine are associated with lower HIV-1 infection risk in an efficacy trial. *The Journal of Infectious Diseases*, 215(9):1376–1385, 2017.

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007.

References IV

Scott D Neidich, Youyi Fong, Shuying S Li, Daniel E Geraghty, Brian D Williamson, William Chad Young, Derrick Goodman, Kelly E Seaton, Xiaoying Shen, Sheetal Sawant, et al. Antibody Fc effector functions and IgG3 associate with decreased HIV-1 risk. *Journal of Clinical Investigation*, 129(11):4838–4849, 2019.

J. Neyman, K. Iwaszkiewicz, and St. Kołodziejczyk. Statistical Problems in Agricultural Experimentation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2(2):107–154, July 1935. ISSN 1369-7412, 1467-9868. doi: 10.2307/2983637. URL https://academic.oup.com/jrsssb/article/2/2/107/7026422. Publisher: Oxford University Press (OUP).

Steven J Novick and Leonard A Stefanski. Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97(458):472–481, June 2002. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214502760047005. URL http://www.tandfonline.com/doi/abs/10.1198/016214502760047005.

David Puelz, Guillaume Basse, Avi Feller, and Panos Toulis. A Graph-Theoretic Approach to Randomization Tests of Causal Effects under General Interference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):174–204, February 2022. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12478. URL https://academic.oup.com/jrsssb/article/84/1/174/7056130.

David M Ritzwoller, Joseph P Romano, and Azeem M Shaikh. Randomization inference: Theory and applications. 2024.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477): 191–200, March 2007. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214506000001112.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1):41–55, 1983. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/70.1.41.

References V

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66 (5):688–701, October 1974. ISSN 1939-2176, 0022-0663. doi: 10.1037/h0037350. URL https://doi.apa.org/doi/10.1037/h0037350. Publisher: American Psychological Association (APA).

Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591, 1980. ISSN 01621459. doi: 10.2307/2287653.

Bonnie E. Shook-Sa, Paul N. Zivich, Stehen R. Cole, Nora E. Rosenberg, Michael G. Hudgens, Deborah J. Donnell, Sizulu Moyo, Khangelani Zuma, and Jessie K. Edwards. Examining the impact of two popart interventions on hiv risk in zambia and south africa. 2025+.

Michael E Sobel. What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, December 2006. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214506000000636.

Leonard A Stefanski and Dennis D Boos. The calculus of M-estimation. *The American Statistician*, 56(1):29–38, 2002.

Yao Zhang and Qingyuan Zhao. What is a randomization test? *Journal of the American Statistical Association*, 118(544):2928–2942, October 2023. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2023.2199814.

Zhiwei Zhang, Jie Zhou, Weihua Cao, and Jun Zhang. Causal inference with a quantitative exposure. *Statistical Methods in Medical Research*, 25(1):315–335, February 2016. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280212452333. URL http://journals.sagepub.com/doi/10.1177/0962280212452333.

References VI

P.N. Zivich, M.C. Eisenberg, A.S. Monto, A. Uzicanin, R. S. Baric, T. P. Sheahan, J. J. Rainey, H. Gao, and A. E. Aiello. Transmission of viral pathogens in a social network of university students: the eX-FLU study. *Epidemiology and Infection*, 148:e267, 2020. ISSN 0950-2688, 1469-4409. doi: 10.1017/S0950268820001806. URL https://www.cambridge.org/core/product/identifier/S0950268820001806/type/journal_article.

Supplement to Project 1 : Addressing Confounding and Continuous Exposure Measurement Error Using Corrected Score Functions

Brian Richardson, Brian Blette, Peter Gilbert, Michael Hudgens

Corrected Score Functions: What?

- Suppose the oracle score function is **conditionally unbiased**, meaning

$$\mathrm{E}\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \mathbf{A}; \theta) | \mathbf{A}\} = \mathbf{0}.$$

- Define the corrected score function as

$$\boldsymbol{\Psi}_{CS}\left(Y, \mathbf{L}, \mathbf{A}^*; \theta\right) = \mathrm{E}\left[\mathrm{Re}\left\{\boldsymbol{\Psi}_0(Y, \mathbf{L}, \widetilde{\mathbf{A}}; \theta)\right\} | Y, \mathbf{L}, \mathbf{A}^*\right],$$

where $\widetilde{\mathbf{A}} = \mathbf{A}^* + i\widetilde{\epsilon}$, $i = \sqrt{-1}$, $\mathrm{Re}(\cdot)$ denotes the real component of a complex number, and $\widetilde{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{me})$.

- Then

$$\begin{aligned}
& \mathrm{E}\left\{\boldsymbol{\Psi}_{CS}\left(Y, \mathbf{L}, \mathbf{A}^*; \theta\right) | Y, \mathbf{L}, \mathbf{A}\right\} = \boldsymbol{\Psi}_0\left(Y, \mathbf{L}, \mathbf{A}; \theta\right) \\
\Longrightarrow\ & \mathrm{E}\left[\mathrm{E}\left\{\boldsymbol{\Psi}_{CS}\left(Y, \mathbf{L}, \mathbf{A}^*; \theta\right) | Y, \mathbf{L}, \mathbf{A}\right\}\right] = \mathrm{E}\left\{\boldsymbol{\Psi}_0\left(Y, \mathbf{L}, \mathbf{A}; \theta\right)\right\} \\
\Longrightarrow\ & \mathrm{E}\left\{\boldsymbol{\Psi}_{CS}\left(Y, \mathbf{L}, \mathbf{A}^*; \theta\right)\right\} = \mathbf{0}
\end{aligned}$$

Corrected Score Functions: Why?

The key result (Novick and Stefanski, 2002) for corrected score functions is that, for a smooth enough function $\mathbf{f}$, the function $\widetilde{\mathbf{f}}$ defined by

$$\widetilde{\mathbf{f}}(\mathbf{A}^*) \equiv \mathrm{E}[\mathrm{Re}\{\mathbf{f}(\mathbf{A}^* + i\widetilde{\epsilon})\}|\mathbf{A}, \mathbf{A}^*]$$

does not depend on $\mathbf{A}$ and satisfies

$$\mathrm{E}\{\widetilde{\mathbf{f}}(\mathbf{A}^*)|\mathbf{A}\} = \mathbf{f}(\mathbf{A}).$$

The proof of this for the special case $f(\mathbf{A}) = \exp\left(\mathbf{c}\mathbf{A}^T\right)$ is illustrative.

Corrected Score Functions: Why?

CLAIM: $\mathrm{E}\{\widetilde{f}(\mathbf{A}^*)|\mathbf{A}\} = f(\mathbf{A})$ for $f(\mathbf{A}) = \exp\left(\mathbf{cA}^T\right)$.

Corrected Score Functions: Why?

CLAIM: $\mathrm{E}\{\widetilde{f}(\mathbf{A}^*)|\mathbf{A}\} = f(\mathbf{A})$ for $f(\mathbf{A}) = \exp\left(\mathbf{c}\mathbf{A}^T\right)$.

$$\widetilde{f}(\mathbf{A}^*) \equiv \mathrm{E}(\mathrm{Re}[\exp\{\mathbf{c}(\mathbf{A}^* + i\widetilde{\epsilon})^T\}]|\mathbf{A}, \mathbf{A}^*)$$
$$= \exp\left(\mathbf{c}\mathbf{A}^{*T}\right)\mathrm{Re}[\underbrace{\mathrm{E}\{\exp\left(i\mathbf{c}\widetilde{\epsilon}^T\right)\}]}_{\text{normal c.f.}} = \exp\left(\mathbf{c}\mathbf{A}^{*T}\right)\exp\left(-\frac{1}{2}\mathbf{c}\mathbf{\Sigma}_{me}\mathbf{c}^T\right)$$

Corrected Score Functions: Why?

CLAIM: $\mathrm{E}\{\widetilde{f}(\mathbf{A}^*)|\mathbf{A}\} = f(\mathbf{A})$ for $f(\mathbf{A}) = \exp\left(\mathbf{c}\mathbf{A}^T\right)$.

$$
\begin{aligned}
\widetilde{f}(\mathbf{A}^*) &\equiv \mathrm{E}(\mathrm{Re}[\exp\{\mathbf{c}(\mathbf{A}^* + i\widetilde{\epsilon})^T\}]|\mathbf{A}, \mathbf{A}^*) \\
&= \exp\left(\mathbf{c}\mathbf{A}^{*T}\right)\mathrm{Re}[\underbrace{\mathrm{E}\{\exp\left(i\mathbf{c}\widetilde{\epsilon}^T\right)\}]}_{\text{normal c.f.}} = \exp\left(\mathbf{c}\mathbf{A}^{*T}\right)\exp\left(-\frac{1}{2}\mathbf{c}\mathbf{\Sigma}_{me}\mathbf{c}^T\right) \\
\Longrightarrow \quad \mathrm{E}\{\widetilde{f}(\mathbf{A}^*)|\mathbf{A}\} &= \mathrm{E}\{\exp\left(\mathbf{c}\mathbf{A}^{*T}\right)|\mathbf{A}\}\exp\left(-\frac{1}{2}\mathbf{c}\mathbf{\Sigma}_{me}\mathbf{c}^T\right) \\
&= \mathrm{E}\{\exp\{\mathbf{c}(\mathbf{A}^T + \epsilon^T)\}|\mathbf{A}\}\exp\left(-\frac{1}{2}\mathbf{c}\mathbf{\Sigma}_{me}\mathbf{c}^T\right) \\
&= \exp\left(\mathbf{c}\mathbf{A}^T\right)\underbrace{\mathrm{E}\{\exp\left(\mathbf{c}\epsilon^T\right)\}}_{\text{normal m.g.f.}}\exp\left(-\frac{1}{2}\mathbf{c}\mathbf{\Sigma}_{me}\mathbf{c}^T\right) \\
&= \exp\left(\mathbf{c}\mathbf{A}^T\right)\exp\left(\frac{1}{2}\mathbf{c}\mathbf{\Sigma}_{me}\mathbf{c}^T\right)\exp\left(-\frac{1}{2}\mathbf{c}\mathbf{\Sigma}_{me}\mathbf{c}^T\right) \\
&= \exp\left(\mathbf{c}\mathbf{A}^T\right) = f(\mathbf{A})
\end{aligned}
$$

Additional Simulation Results: G-Formula

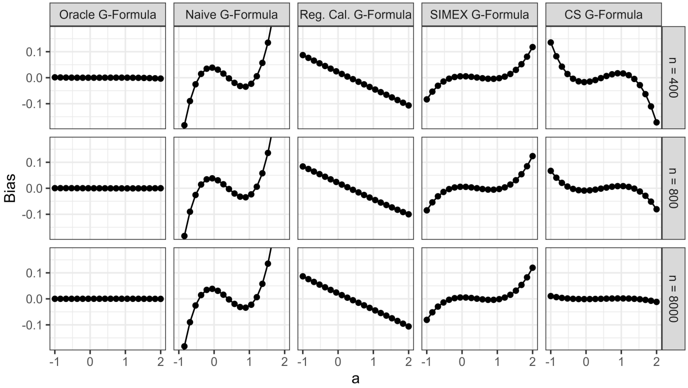


Figure: Estimated dose-response curve bias for the oracle, naive, regression calibration, SIMEX, and CS g-formula estimators. Bias refers to the average bias across 1000 simulated data sets for each method evaluated at each point on the horizontal axis, corresponding to setting the true exposure to $a \in [-1, 2]$. For the naive estimator, biases outside of $[-0.18, 0.18]$ are excluded from the plot.

Additional Simulation Results: IPW

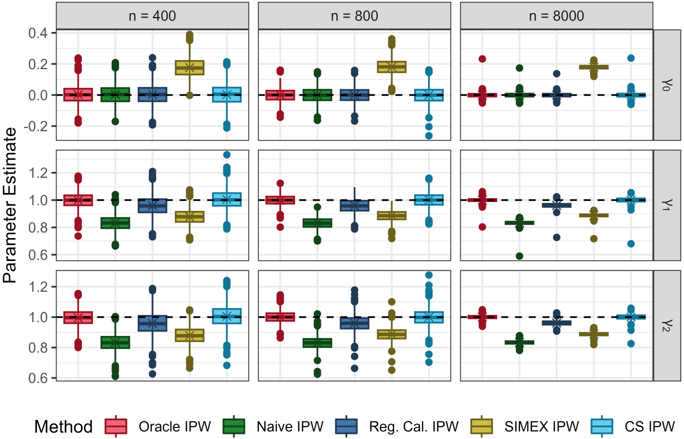


Figure: Empirical distribution of MSM parameter estimates from the simulation study of IPW estimators, using the oracle, naive, regression calibration (reg. cal.), simulation-extrapolation (SIMEX), and corrected score (CS) IPW estimators.

Additional Simulation Results: Doubly Robust Estimator

Table: Results from the simulation study of DR estimators. UC: uncorrected empirical sandwich variance estimator, BC: bias-corrected empirical sandwich variance estimator. PS indicates the propensity score model is correctly specified; OR indicates the outcome regression is correctly specified.

					UC		BC	
n	**Correct Specifications**	**Method**	**Bias**	**ESE**	**ASE**	**Cov**	**ASE**	**Cov**
2000	PS and OR	CS DR	0.2	8.5	7.9	92.6	7.9	92.9
		CS G-Formula	0.0	6.8	6.6	94.2	6.6	94.3
		CS IPW	0.1	8.9	8.3	93.5	8.4	94.0
	PS Only	CS DR	-0.1	8.6	8.0	93.3	8.1	93.4
		CS G-Formula	-14.9	6.7	6.6	39.7	6.6	40.0
		CS IPW	0.1	8.9	8.3	93.5	8.4	94.0
	OR Only	CS DR	0.1	7.7	7.2	93.0	7.2	93.0
		CS G-Formula	0.0	6.8	6.6	94.2	6.6	94.3
		CS IPW	-14.7	7.6	7.3	47.9	7.3	47.9
	Neither	CS DR	-14.7	7.5	7.1	47.0	7.2	47.3
		CS G-Formula	-14.9	6.7	6.6	39.7	6.6	40.0
		CS IPW	-14.7	7.6	7.3	47.9	7.3	47.9

Additional Simulation Results: Estimated Measurement Error Variance

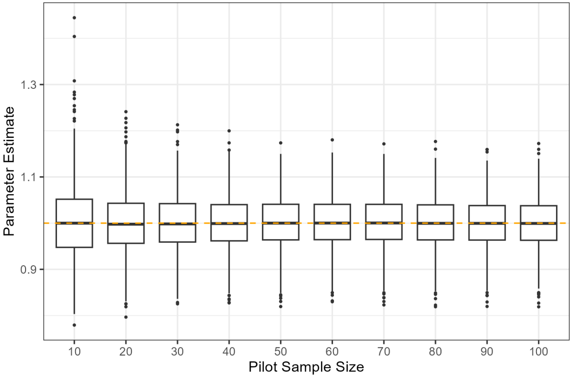


Figure: Empirical distribution of estimated γ_1 versus pilot study sample size n_p, using the corrected score (CS) IPW estimator with estimated measurement error covariance.

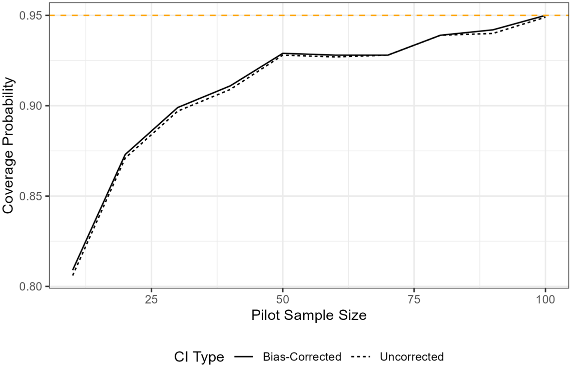


Figure: Empirical coverage probabilities of 95% confidence intervals for γ_1 versus pilot study sample size n_p, using the corrected score (CS) IPW estimator with estimated measurement error covariance.

Additional Simulation Results: Near-Positivity Violation

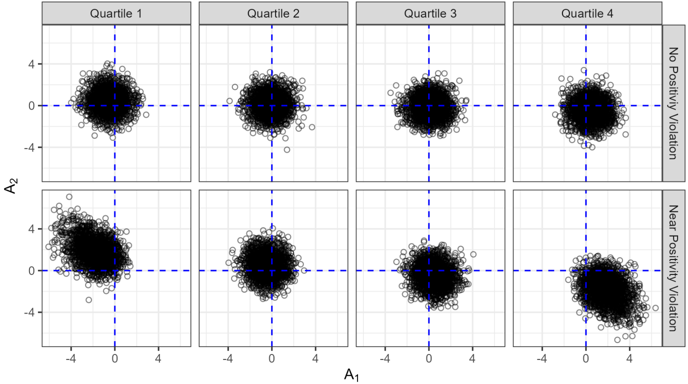


Figure: True exposure values (A_1, A_2) by quartile of L from one simulated data set of sample size $n = 8000$, with and without a near positivity violation.

Additional Simulation Results: Near-Positivity Violation

Table: Results from the simulation study with a near positivity violation. UC: uncorrected empirical sandwich variance estimator, BC: bias-corrected empirical sandwich variance estimator, n: sample size; Bias: 100 times the average bias across simulated data sets for each method; ESE: 100 times the standard deviation of parameter estimates; ASE: 100 times the average of estimated standard errors; Cov: Empirical percent coverage of 95% confidence intervals for each method.

					UC		BC	
n	**Method**	**Parameter**	**Bias**	**ESE**	**ASE**	**Cov**	**ASE**	**Cov**
8000	Oracle IPW	γ_0	-0.8	10.8	6.4	90.4	6.6	91.0
		γ_1	4.2	8.2	4.5	65.6	5.2	69.9
		γ_2	-3.6	8.2	4.7	64.3	5.4	68.7
	Naive IPW	γ_0	-0.4	10.7	6.8	89.1	7.0	89.5
		γ_1	-13.0	7.9	4.6	17.9	5.1	21.6
		γ_2	-19.9	8.0	4.6	10.2	5.1	12.2
	CS IPW	γ_0	-2.3	46.0	44.9	89.9	74.3	90.1
		γ_1	4.4	66.4	61.9	68.0	106.0	71.5
		γ_2	-4.3	33.0	31.0	76.6	49.8	78.8

Additional Simulation Results: Multiplicative Measurement Error

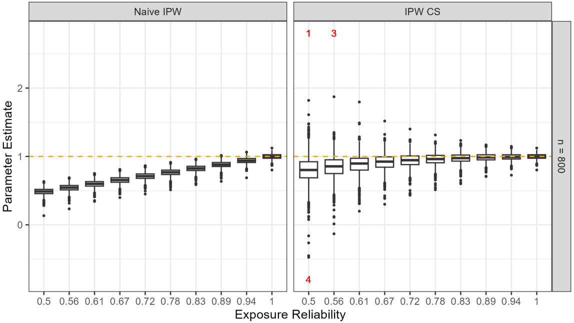


Figure: Empirical distribution of estimated γ_1 versus exposure reliability $\mathrm{Var}(A_1)/\mathrm{Var}(A_1^*) = \mathrm{Var}(A_2)/\mathrm{Var}(A_2^*)$, using naive and corrected score (CS) IPW estimators and for a multiplicative measurement error data generating process. For the CS IPW estimator, the number values beyond the ranges of the plot are shown in red on the tails of the boxplots.

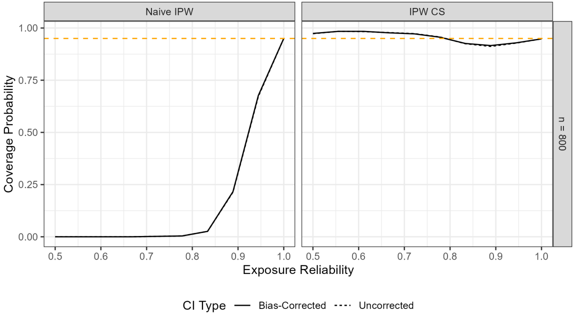


Figure: Empirical coverage probabilities of 95% confidence intervals for γ_1 versus exposure reliability $\mathrm{Var}(A_1)/\mathrm{Var}(A_1^*) = \mathrm{Var}(A_2)/\mathrm{Var}(A_2^*)$, using naive and corrected score (CS) IPW estimators for a multiplicative measurement error data generating process.

Supplement to Project 2: Social Distancing to Reduce Transmission of Influenza-Like-Illness on College Campuses: the eX-FLU Trial

Brian Richardson, Allison Aiello, Michael Hudgens

Choice of Test Statistic

Test statistics: "proportion of possible transmission events attributable to students in the intervention group:"

$$T^{**}(\mathbf{Z}, \overline{\mathbf{A}}, \overline{\mathbf{Y}}) = \frac{\sum_{k=2}^{\tau} \sum_{i=1}^{n} \sum_{j>i} Z_i E_{ijk}^{**}}{\sum_{k=2}^{\tau} \sum_{i=1}^{n} \sum_{j>i} E_{ijk}^{**}}.$$

	From Infected to Infected	**From Infected**
Contact at $k-1$	$E_{ijk}^{11} = Y_i^{k-1} A_{ij}^{k-1} Y_j^k$	$E_{ijk}^{12} = Y_i^{k-1} A_{ij}^{k-1}$
Contact at k	$E_{ijk}^{21} = Y_i^{k-1} A_{ij}^{k} Y_j^k$	$E_{ijk}^{22} = Y_i^{k-1} A_{ij}^{k}$
Contact at $k-1$ **or** k	$E_{ijk}^{31} = Y_i^{k-1} (A_{ij}^{k-1} \vee A_{ij}^{k}) Y_j^k$	$E_{ijk}^{32} = Y_i^{k-1} (A_{ij}^{k-1} \vee A_{ij}^{k})$
Contact at $k-1$ **and** k	$E_{ijk}^{41} = Y_i^{k-1} (A_{ij}^{k-1} * A_{ij}^{k}) Y_j^k$	$E_{ijk}^{42} = Y_i^{k-1} (A_{ij}^{k-1} * A_{ij}^{k})$

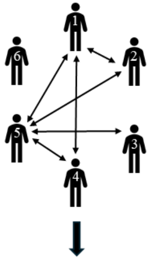
Table: Definitions of a possible transmission event E_{ijk} from student i to student j at time k. Y_i^k is an indicator for student i being infected at week k, A_{ij}^k is an indicator for students i and j being in contact at week k, and $a \vee b$ denotes the maximum of a and b.

Exponential Family Random Graph Models

Goal: model a network ($\mathbf{A}$) given covariates ($\mathbf{X}$)

Exponential Family Random Graph Models

Goal: model a network (**A**) given covariates (**X**)



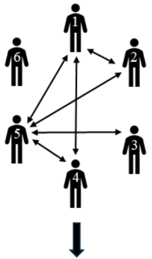
	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

Exponential Family Random Graph Models

Goal: model a network ($\mathbf{A}$) given covariates ($\mathbf{X}$)

An **ERGM** assumes:

$$\Pr_{\theta}(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}}{\kappa(\theta, \mathcal{A}, \mathbf{x})}$$



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

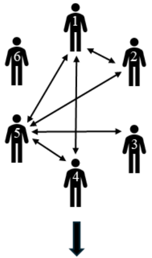
Exponential Family Random Graph Models

Goal: model a network (**A**) given covariates (**X**)

An **ERGM** assumes:

$$\Pr_{\theta}(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}}{\kappa(\theta, \mathcal{A}, \mathbf{x})}$$

- normalizing constant:
  $\kappa(\theta, \mathcal{A}, \mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}} \exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}$



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

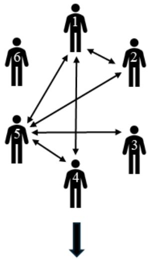
Exponential Family Random Graph Models

Goal: model a network ($\mathbf{A}$) given covariates ($\mathbf{X}$)

An **ERGM** assumes:

$$\Pr_{\theta}(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}}{\kappa(\theta, \mathcal{A}, \mathbf{x})}$$

- normalizing constant:
  $\kappa(\theta, \mathcal{A}, \mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}} \exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}$
- $\mathcal{A}$: space of possible networks



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

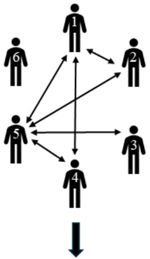
Exponential Family Random Graph Models

Goal: model a network (**A**) given covariates (**X**)

An **ERGM** assumes:

$$\Pr_{\theta}(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}}{\kappa(\theta, \mathcal{A}, \mathbf{x})}$$

- normalizing constant: $\kappa(\theta, \mathcal{A}, \mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}} \exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}$
- $\mathcal{A}$: space of possible networks
- θ: model coefficients



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

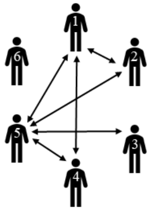
Exponential Family Random Graph Models

Goal: model a network (**A**) given covariates (**X**)

An **ERGM** assumes:

$$\Pr_{\theta}(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}}{\kappa(\theta, \mathcal{A}, \mathbf{x})}$$

- normalizing constant:
  $\kappa(\theta, \mathcal{A}, \mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}} \exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}$
- $\mathcal{A}$: space of possible networks
- θ: model coefficients
- $\mathbf{g}(\mathbf{a}, \mathbf{x})$: sufficient statistics



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

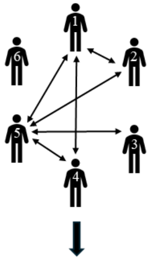
Exponential Family Random Graph Models

Goal: model a network (**A**) given covariates (**X**)

An **ERGM** assumes:

$$\Pr_{\theta}(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}}{\kappa(\theta, \mathcal{A}, \mathbf{x})}$$

- normalizing constant: $\kappa(\theta, \mathcal{A}, \mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}} \exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}$
- $\mathcal{A}$: space of possible networks
- θ: model coefficients
- $\mathbf{g}(\mathbf{a}, \mathbf{x})$: sufficient statistics
  - number of edges



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

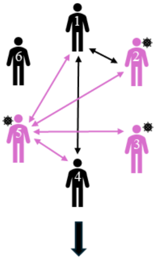
Exponential Family Random Graph Models

Goal: model a network ($\mathbf{A}$) given covariates ($\mathbf{X}$)

An **ERGM** assumes:

$$\Pr_\theta(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}}{\kappa(\theta, \mathcal{A}, \mathbf{x})}$$

- normalizing constant:
  $\kappa(\theta, \mathcal{A}, \mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}} \exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}$
- $\mathcal{A}$: space of possible networks
- θ: model coefficients
- $\mathbf{g}(\mathbf{a}, \mathbf{x})$: sufficient statistics
  - number of edges
  - number of edges touching treated students



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

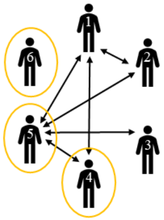
Exponential Family Random Graph Models

Goal: model a network ($\mathbf{A}$) given covariates ($\mathbf{X}$)

An **ERGM** assumes:

$$\Pr_{\theta}(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}}{\kappa(\theta, \mathcal{A}, \mathbf{x})}$$

- normalizing constant: $\kappa(\theta, \mathcal{A}, \mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}} \exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}$
- $\mathcal{A}$: space of possible networks
- θ: model coefficients
- $\mathbf{g}(\mathbf{a}, \mathbf{x})$: sufficient statistics
  - number of edges
  - number of edges touching treated students
  - number of edges touching infected students



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

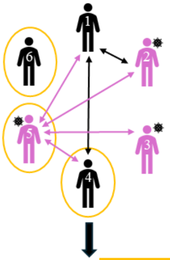
Exponential Family Random Graph Models

Goal: model a network ($\mathbf{A}$) given covariates ($\mathbf{X}$)

An **ERGM** assumes:

$$\Pr_{\theta}(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}}{\kappa(\theta, \mathcal{A}, \mathbf{x})}$$

- normalizing constant:
  $\kappa(\theta, \mathcal{A}, \mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}} \exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}$
- $\mathcal{A}$: space of possible networks
- θ: model coefficients
- $\mathbf{g}(\mathbf{a}, \mathbf{x})$: sufficient statistics
  - number of edges
  - number of edges touching treated students
  - number of edges touching infected students
  - number of edges touching treated $\times$ infected students



	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	0	0	1	0
3	0	0	0	0	1	0
4	1	0	0	0	1	0
5	1	1	1	1	0	0
6	0	0	0	0	0	0

ERGM Model Formulation

$$\Pr_{\theta}(\mathbf{A} = \mathbf{a} | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}}{\kappa(\theta, \mathcal{A}, \mathbf{x})},$$

$$\kappa(\theta, \mathcal{A}, \mathbf{x}) = \sum_{\mathbf{a} \in \mathcal{A}} \exp\{\theta \cdot \mathbf{g}(\mathbf{a}, \mathbf{x})\}$$

For example, in the simulation study and eX-FLU application,

$$\mathbf{g}(\mathbf{a}, \mathbf{x}) = \begin{bmatrix} \text{\# edges} \\ \text{\# edges touching a treated node} \\ \text{\# edges touching an infected node} \\ \text{\# edges touching a treated and infected node} \\ \text{\# edges between roommate pairs} \end{bmatrix} = \begin{bmatrix} \sum_{i,j} a_{ij} \\ \sum_{i,j} a_{ij}(z_i + z_j) \\ \sum_{i,j} a_{ij}(y_i + y_j) \\ \sum_{i,j} a_{ij}(z_i y_i + z_j y_j) \\ \sum_{i,j} a_{ij} \mathbb{1}(i, j \text{ roomates}) \end{bmatrix}$$

ERGM Change Statistic Model Formulation

- **Change statistic**: $\delta_{\mathbf{g}}(\mathbf{a}, \mathbf{x})_{ij} = \mathbf{g}(\mathbf{a}_{ij}^{+}, \mathbf{x}) - \mathbf{g}(\mathbf{a}_{ij}^{-}, \mathbf{x})$ is the change in network statistic that would occur if a_{ij} were changed from 0 to 1
  - where $\mathbf{a}_{ij}^{+}$ and $\mathbf{a}_{ij}^{-}$ represent the network $\mathbf{a}$ with dyad a_{ij} set to 1 or 0, respectively
- Then the equivalent ERGM specification is

$$\text{logit}\{\Pr_{\theta}(A_{ij} = 1 | \mathbf{A}_{ij}^{C} = \mathbf{a}_{ij}^{C}, \mathbf{X} = \mathbf{x})\} = \theta^{k} \delta_{\mathbf{g}}(\mathbf{a}, \mathbf{x})_{ij}$$

  - where $\mathbf{A}_{ij}^{C}$ represents all dyads in $\mathbf{A}$ except A_{ij}
- **Interpretation of θ**: the change in conditional log-odds of the network associated with a one-unit increase in the corresponding component of $\mathbf{g}(\mathbf{a}, \mathbf{x})$ resulting from switching a particular dyad A_{ij} from 0 to 1 and leaving the rest of the network fixed at $\mathbf{A}_{ij}^{C}$

Dyadic Independence EGRMs

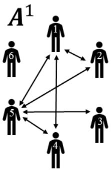
- **Dyadic independence term**: a component g of $\mathbf{g}$ in an ERGM for which the corresponding change statistic $\delta_g(\mathbf{a}, \mathbf{x})_{ij}$ can be calculated for any i, j without knowing $\mathbf{a}$
  - For example, if $g(\mathbf{a}, \mathbf{x}) = \sum_{i,j} a_{ij}(Z_i + Z_j)$ counts the number of edges touching treated nodes, then $\delta_g(\mathbf{a}, \mathbf{x})_{ij} = z_i + z_j$ doesn't depend on $\mathbf{a}$
- **Dyadic independence ERGM**: an ERGM with only dyadic independence terms
  - replace $\delta_g(\mathbf{a}, \mathbf{x})_{ij}$ with $\delta_g(\mathbf{x})_{ij}$ and write the model as

$$\text{logit}\{\Pr\big(A_{ij} = 1 | \mathbf{X} = \mathbf{x}\big)\} = \theta \cdot \delta_g(\mathbf{x})_{ij}$$

- **Interpretation of θ**: the change in log-odds of the network associated with a one-unit increase in the corresponding component of $\mathbf{g}(\mathbf{a}, \mathbf{x})$ resulting from switching a particular dyad A_{ij} from 0 to 1

Separable Temporal Exponential Family Random Graph Models

A **STERGM** assumes that, at each time step $k = 1, \ldots, \tau - 1$:

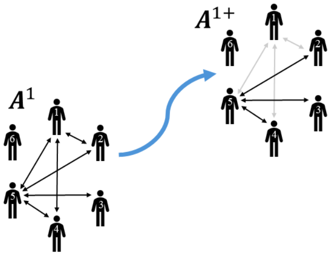


Separable Temporal Exponential Family Random Graph Models

A **STERGM** assumes that, at each time step $k = 1, \ldots, \tau - 1$:

- New edges form according to a **formation** ERGM

$$\Pr_{\theta^+}(\mathbf{A}^{k+} = \mathbf{a}^{k+} | \mathbf{A}^k = \mathbf{a}^k, \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta^+ \cdot \mathbf{g}^+(\mathbf{a}^{k+}, \mathbf{x})\}}{\kappa\{\theta^+, \mathcal{A}^+(\mathbf{a}^k), \mathbf{x}\}}$$

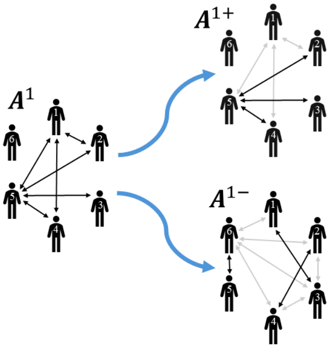


Separable Temporal Exponential Family Random Graph Models

A **STERGM** assumes that, at each time step $k = 1, \dots, \tau - 1$:

- New edges form according to a **formation** ERGM
- Old edges persist according to a **persistence** ERGM

$$\Pr_{\theta^-}(\mathbf{A}^{k-} = \mathbf{a}^{k-} | \mathbf{A}^k = \mathbf{a}^k, \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta^- \cdot \mathbf{g}^-(\mathbf{a}^{k-}, \mathbf{x})\}}{\kappa\{\theta^-, \mathcal{A}^-(\mathbf{a}^k), \mathbf{x}\}}.$$

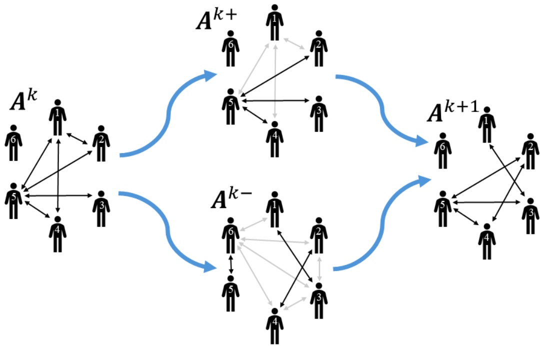


Separable Temporal Exponential Family Random Graph Models

A **STERGM** assumes that, at each time step $k = 1, \ldots, \tau - 1$:

- New edges form according to a **formation** ERGM
- Old edges persist according to a **persistence** ERGM
- The network at time step $k + 1$ is the result of formation and persistence

$$\mathbf{A}^{k+1} = \underbrace{\mathbf{A}^{k}}_{\text{previous network}} \cup \underbrace{(\mathbf{A}^{k+} - \mathbf{A}^{k})}_{\text{new edges formed}} - \underbrace{(\mathbf{A}^{k} - \mathbf{A}^{k-})}_{\text{old edges not persisting}}$$



STERGM Model Formulation

- **Formation model** is an ERGM conditional on only adding edges:

$$\Pr_{\theta^+}(\mathbf{A}^{k+} = \mathbf{a}^{k+} | \mathbf{A}^k = \mathbf{a}^k, \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta^+ \cdot \mathbf{g}^+(\mathbf{a}^{k+}, \mathbf{x})\}}{\kappa\{\theta^+, \mathcal{A}^+(\mathbf{a}^k), \mathbf{x}\}}$$

  - $\mathcal{A}^+(\mathbf{a})$: space of possible networks that can be formed by adding edges to $\mathbf{a}$

- **Persistence model** is an ERGM conditional on only removing edges:

$$\Pr_{\theta^-}(\mathbf{A}^{k-} = \mathbf{a}^{k-} | \mathbf{A}^k = \mathbf{a}^k, \mathbf{X} = \mathbf{x}) = \frac{\exp\{\theta^- \cdot \mathbf{g}^-(\mathbf{a}^{k-}, \mathbf{x})\}}{\kappa\{\theta^-, \mathcal{A}^-(\mathbf{a}^k), \mathbf{x}\}}$$

  - $\mathcal{A}^-(\mathbf{a})$: space of possible networks that can be formed by removing edges to $\mathbf{a}$

- A **STERGM** assumes the network at time $k+1$ is then the result of applying the changes in $\mathbf{A}^{k+}$ and $\mathbf{A}^{k-}$ to $\mathbf{A}^k$:

$$\mathbf{A}^{k+1} = \underbrace{\mathbf{A}^k}_{\text{previous network}} \cup \underbrace{(\mathbf{A}^{k+} - \mathbf{A}^k)}_{\text{new edges formed}} - \underbrace{(\mathbf{A}^k - \mathbf{A}^{k-})}_{\text{old edges not persisting}}$$

Separability of STERGMs

1. $\mathbf{A}^{k+} \perp\!\!\!\perp \mathbf{A}^{k-} | \mathbf{A}^{k}$, i.e., the formation and persistence processes are conditionally independent give the network at time k

Separability of STERGMs

1. $\mathbf{A}^{k+} \perp\!\!\!\perp \mathbf{A}^{k-} | \mathbf{A}^k$, i.e., the formation and persistence processes are conditionally independent give the network at time k

$$\begin{tikzcd}
& \mathbf{A}^{1+} \arrow[dr] & & \mathbf{A}^{2+} \arrow[dr] & & \mathbf{A}^{3+} \arrow[dr] & \\
\mathbf{A}^1 \arrow[ur] \arrow[rr] \arrow[dr] & & \mathbf{A}^2 \arrow[ur] \arrow[rr] \arrow[dr] & & \mathbf{A}^3 \arrow[ur] \arrow[rr] \arrow[dr] & & \mathbf{A}^4 \\
& \mathbf{A}^{1-} \arrow[ur] & & \mathbf{A}^{2-} \arrow[ur] & & \mathbf{A}^{3-} \arrow[ur] &
\end{tikzcd}$$

Separability of STERGMs

1. $\mathbf{A}^{k+} \perp\!\!\!\perp \mathbf{A}^{k-} | \mathbf{A}^k$, i.e., the formation and persistence processes are conditionally independent give the network at time k
2. the parameter space for $\theta = (\theta^+, \theta^-)$ is the product of the parameter spaces for θ^+ and θ^-

$$
\begin{tikzcd}
 & \mathbf{A}^{1+} \arrow[dr] & & \mathbf{A}^{2+} \arrow[dr] & & \mathbf{A}^{3+} \arrow[dr] & \\
\mathbf{A}^1 \arrow[ur] \arrow[rr] \arrow[dr] & & \mathbf{A}^2 \arrow[ur] \arrow[rr] \arrow[dr] & & \mathbf{A}^3 \arrow[ur] \arrow[rr] \arrow[dr] & & \mathbf{A}^4 \\
 & \mathbf{A}^{1-} \arrow[ur] & & \mathbf{A}^{2-} \arrow[ur] & & \mathbf{A}^{3-} \arrow[ur] &
\end{tikzcd}
$$

Simulation Setup

- $n \in \{24, 48\}$ students were equally divided into two residence halls, each with six clusters of equal size
- Within each residence hall group, the clusters were randomized using a 50:50 allocation to either the intervention group or the control group
- Five pairs of students were randomly selected to be roommates, meaning they had contact with each other each week
- Baseline $(k = 1)$ social contacts were simulated between each pair of students with probability 0.5
- Baseline infection statuses were simulated for each student with probability 0.5
- Social networks were simulated over the remaining $\tau \in \{5, 10\}$ weeks according to a STERGM with both formation and persistence models including edge count, intervention assignment Z, infection status Y, $Z \times Y$ interaction, and an offset to force constant edges between roommates
- Infection statuses were simulated for each student i at each week $k \in \{2, \ldots, \tau\}$ with probability $\Pr\left(Y_i^k = 1 | \overline{\mathbf{A}}^{k-1}, \overline{\mathbf{Y}}^{k-1}\right) = g(\sum_{j=1}^n A_{ij}^{k-1} Y_j^{k-1})$, where $\sum_{j=1}^n A_{ij}^{k-1} Y_j^{k-1}$ is the number of infected contacts at the previous week, and $g : [0, \infty) \to [0, 1]$ is a non-decreasing function

Simulation Setup

Three scenarios:

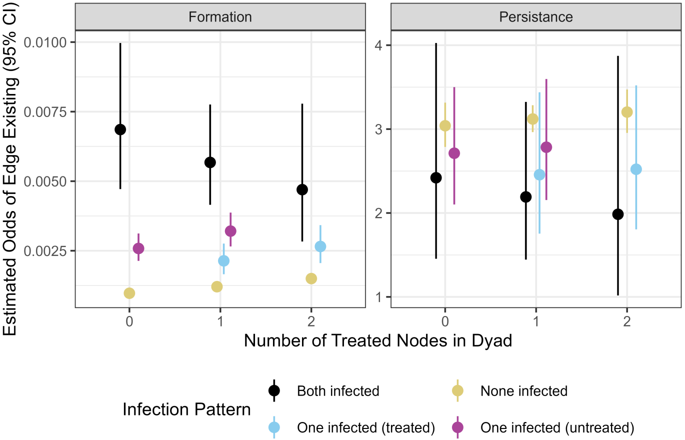
Null Hypothesis	Formation Model Parameters θ^+	Persistence Model Parameters θ^-	Infection Probability Function g
$H_0^\sharp$	$(-0.5, 0, 0, 0)$	$(-0.5, 0, 0, 0)$	$g(s) = 0.5$
$\overline{H}_0^A \cap H_0^Y$	$(-0.2, 0, 0, -1)$	$(-0.2, 0, 0, -1)$	$g(s) = 0.5$
$\overline{H}_0^A \cap \overline{H}_0^Y$	$(-0.2, 0, 0, -1)$	$(-0.2, 0, 0, -1)$	$g(s)$ increasing in s

Table: Data generating process for the simulation study. The null hypothesis refers to the null hypothesis that is true under the data generating process, formation model parameters are $\theta^+ = (\theta^+_{\text{edges}}, \theta^+_Z, \theta^+_Y, \theta^+_{ZY})$, persistence model parameters are $\theta^- = (\theta^-_{\text{edges}}, \theta^-_Z, \theta^-_Y, \theta^-_{ZY})$, and the infection probability function g provides a unit's probability of being infected at week k given their number of infected neighbors at week $k-1$.

Three p-values:

- $\rho_B^\sharp$: testing the sharp null
- ρ_B^Y: testing H_0^Y using known q
- $\widehat{\rho}_B^Y$: testing H_0^Y using estimated q

eX-FLU: STERGM Results



Type I Error Control

- **Proposition 2.1**: Let T_N be a test statistic with CDF F_N under H_0.
  1. Under H_0, $F_N(T_N)$ stochastically dominates a Uniform$(0,1)$ distribution for any N.
  2. If the test statistic T_N has a continuous limiting distribution, then $F_N(T_N) \to^d$ Uniform$(0,1)$.
- **Corollary 2.2**:
  1. Under $H_0^\sharp$, the sharp null p-value $\rho_N^\sharp$ stochastically dominates a Uniform$(0,1)$ distribution for any N.
  2. Under H_0^Y, the oracle p-value ρ_N^Y stochastically dominates a Uniform$(0,1)$ distribution for any N.
  3. If the test statistic T_N has a continuous limiting distribution, then $\rho_N^\sharp \to^d$ Uniform$(0,1)$ under $H_0^\sharp$ and $\rho_N^Y \to^d$ Uniform$(0,1)$ under H_0^Y.

Type I Error Control

- **Proposition 2.3**: Let $q(\mathbf{a}, \mathbf{z}, \theta) \equiv \Pr\{\mathbf{A}(\mathbf{z}) = \mathbf{a}; \theta\}$ denote the PMF of the distribution of stochastic potential networks $\mathbf{A}(\mathbf{z})$ at parameter value θ, let $\widehat{\theta}_N$ denote the estimator of θ, and let $F_N(\cdot; \theta)$ denote the CDF of the test statistic T_N at θ, with limiting CDF $F(\cdot; \theta)$. Let θ_0 denote the true value of θ. Assume the following:

  (A1) $\widehat{\theta}_N \to^p \theta_0$

  (A2) $F(t; \theta_0)$ is continuous in t on $\mathbb{R}$

  (A3) there exists a $\delta_0 > 0$ such that

  $$\sup_{\theta \in B_{\delta_0}(\theta_0)} \sup_{t \in \mathbb{R}} |F_N(t; \theta) - F(t; \theta)| \to 0$$

  (A4) $F(t; \theta)$ is continuous in θ at θ_0 uniformly in t, i.e.,

  $$\lim_{\theta \to \theta_0} \sup_{t \in \mathbb{R}} |F(t; \theta) - F(t; \theta_0)| = 0$$

  Then the plug-in p-value $\widehat{\rho}_N^Y$ converges in distribution to Uniform$(0, 1)$.

Type I Error Control

- **Proposition 2.4**: Let $\rho_N = F_N(T_N)$ for test statistic T_N and CDF F_N (not necessarily the true CDF of T_N). Let $T_N^* = h_N(T_N)$ for a sequence of deterministic, strictly increasing functions h_N. Define $F_N^*(t) = F_N\{h_N^{-1}(t)\}$, the (not necessarily true) CDF of the transformed test statistic, and let $\rho_N^* = F_N^*(T_N^*)$. Then $\rho_N^* = \rho_N$.
- **Corollary 2.5**: Let h_N be a sequence of deterministic, strictly increasing functions.
  1. If the hypotheses of Proposition 2.1 are met for a test statistic T_N, then the results also hold for $T_N^* = h_N(T_N)$.
  2. If the hypotheses of Proposition 2.3 are met for $T_N, F_N(\cdot;\theta)$, then the results also hold for $T_N^* = h_N(T_N), F_N^*(\cdot;\theta) = F_N\{h_N^{-1}(\cdot);\theta\}$.

Supplement to Project 3 : Causal Inference from Cluster Randomized Trials with Differential Nonresponse

Brian Richardson, Bonnie Shook-Sa, Michael Hudgens

Causal Estimand

- **individual-level potential outcomes**:

$$\mathbf{Y}_j(a) = [Y_{j1}(a), \ldots, Y_{jN_j}(a)]$$

Causal Estimand

- **individual-level potential outcomes**:

$$\mathbf{Y}_j(a) = [Y_{j1}(a), \ldots, Y_{jN_j}(a)]$$

- **cluster-level mean potential outcome**:

$$\overline{Y}_j(a) = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ji}(a)$$

Causal Estimand

- **individual-level potential outcomes**:

$$\mathbf{Y}_j(a) = [Y_{j1}(a), \ldots, Y_{jN_j}(a)]$$

- **cluster-level mean potential outcome**:

$$\overline{Y}_j(a) = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ji}(a)$$

- Two ways to aggregate over the population:

Causal Estimand

- **individual-level potential outcomes**:

$$\mathbf{Y}_j(a) = [Y_{j1}(a), \ldots, Y_{jN_j}(a)]$$

- **cluster-level mean potential outcome**:

$$\overline{Y}_j(a) = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ji}(a)$$

- Two ways to aggregate over the population:
  1. Equal weight to each cluster:

$$\eta(a) = \frac{1}{m} \sum_{j=1}^{m} \mathrm{E}\{\overline{Y}_j(a)\},$$

Causal Estimand

- **individual-level potential outcomes**:

$$\mathbf{Y}_j(a) = [Y_{j1}(a), \ldots, Y_{jN_j}(a)]$$

- **cluster-level mean potential outcome**:

$$\overline{Y}_j(a) = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ji}(a)$$

- Two ways to aggregate over the population:
  1. Equal weight to each cluster:

$$\eta(a) = \frac{1}{m} \sum_{j=1}^{m} \mathrm{E}\{\overline{Y}_j(a)\},$$

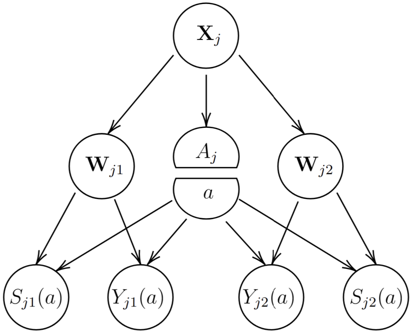
  2. Equal weight to each individual:

$$\eta'(a) = \frac{\sum_{j=1}^{m} N_j \mathrm{E}\{\overline{Y}_j(a)\}}{\sum_{j=1}^{m} N_j}.$$

Working Assumptions for Identifiability

For all clusters j, j', individuals i and treatment a:

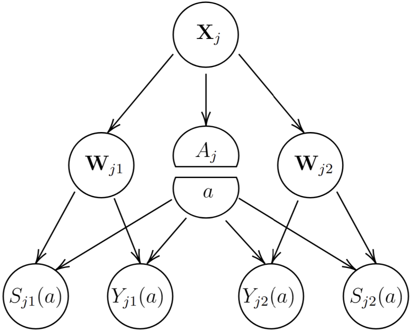
- causal consistency: $Y_{ji} = Y_{ji}(a)$ and $S_{ji} = S_{ji}(a)$ when $A_j = a$
- positivity of treatment assignment:
  $\Pr(A_j = a | \mathbf{X}_j = \mathbf{x}) > 0$ for all $\mathbf{x}$ with $\Pr(\mathbf{X}_j = \mathbf{x}) > 0$
- positivity of response: $\Pr(S_{ji} = 1 | A_j = a, \mathbf{X}_j = \mathbf{x}, \mathbf{W}_{ji} = \mathbf{w}) > 0$
  for all $\mathbf{x}, \mathbf{w}$ with
  $\Pr(\mathbf{X}_j = \mathbf{x}, \mathbf{W}_{ji} = \mathbf{w}) > 0$
- independent clusters:
  $$\left[A_j, \mathbf{X}_j, \{R_{ji}, S_{ji}(a)\}_{i=1}^{N_j}, \mathbf{Y}_j(a), Q_{ji}\right] \perp\!\!\!\perp \left[A_{j'}, \mathbf{X}_{j'}, \{R_{j'i}, S_{j'i}(a)\}_{i=1}^{N_{j'}}, \mathbf{Y}_{j'}(a), Q_{j'i}\right]$$



Working Assumptions for Identifiability

For all clusters j, j', individuals i and treatment a:

- conditional exchangeability: $Y_{ji}(a) \perp\!\!\!\perp A_j | \mathbf{X}_j, \mathbf{W}_{j,i}$
- conditional independence of potential outcome and responses: $Y_{ji}(a) \perp\!\!\!\perp S_{j,i}(a) | A_j, \mathbf{X}_j, \mathbf{W}_{j,i}$
- cluster-level randomization: $A_j \perp\!\!\!\perp (\mathbf{W}_{ji}, \ldots, \mathbf{W}_{jN_j}) | \mathbf{X}_j$
- independent random sampling: $R_{ji} \perp\!\!\!\perp \left[A_j, \mathbf{X}_j, \{S_{ji}(a)\}_{i=1}^{N_j}, \mathbf{Y}_j(a), Q_{ji}\right]$
- independent random sampling of auxiliary data: $Q_{ji} \perp\!\!\!\perp, \left[A_j, \mathbf{X}_j, \{R_{ji}, S_{ji}(a)\}_{i=1}^{N_j}, \mathbf{Y}_j(a)\right]$



Estimators: G-Formula

Define the conditional mean outcome among trial responders with cluster-level covariates $\mathbf{x}$ and individual-level covariates $\mathbf{w}$, and under treatment assignment a by

$$\mu_a(\mathbf{x}, \mathbf{w}) = \mathrm{E}\left(Y_{ji} | A_j = a, R_{ji} = 1, S_{ji} = 1, \mathbf{X}_j = \mathbf{x}, \mathbf{W}_{ji} = \mathbf{w}\right),$$

The g-formula estimator averages $\mu_a(\mathbf{x}, \mathbf{w})$ over the empirical distribution of $(\mathbf{X}_j, \mathbf{W}_{ji})$ in the auxiliary data:

$$\widehat{\eta}^{GF}(a) = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{\widetilde{n}_j} \sum_{i=1}^{N_j} Q_{ji} \mu_a(\mathbf{X}_j, \mathbf{W}_{ji}).$$

Estimators: Inverse Probability Weighting

Define the conditional probability of responding and having treatment assignment to a given covariates.

$$\pi_a(\mathbf{x}, \mathbf{w}) = \Pr(S_{ji} = 1, A_j = a | \mathbf{X}_j = \mathbf{x}, \mathbf{W}_{ji} = \mathbf{w})$$

The IPW estimator weights the observed outcomes of responders using the inverse of $\pi_a(\mathbf{x}, \mathbf{w})$ to create a pseudopopulation that is free of both bias from nonresponse and confounding induced by the conditional randomization design:

$$\widehat{\eta}^{IPW}(a) = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n_j} \sum_{i=1}^{N_j} \frac{\mathbb{1}(A_j = a) R_{ji} S_{ji} Y_{ji}}{\pi_a(\mathbf{X}_j, \mathbf{W}_{ji})}.$$