Missing Covariates in Meta-Regression: Considerations and Implications

Terri Pigott, Georgia State University
Karina Diaz, University of Pennsylvania
Jihyun Lee, University of Texas at Austin
Jacob Schauer, Northwestern University

Table of contents

01

Introduction

Missing Data in Meta-Analysis

Missing data occurs in all meta-analyses, particularly in the form of missing information about potential moderators of effect size heterogeneity

But what to do about this problem?

In the past, we have used complete-case analysis or shifting-case-analysis (available case analysis)

Missing Data Methods and Meta-Analysis

Methods for missing or incomplete data analysis focus on understanding the mechanisms for missing data - what data are missing and why

The appropriate method for analysis when missing data occurs depends on the missing data mechanism

Three major missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random

How do these mechanisms translate to the context of meta-analysis and applications of meta-regression?

Meta-Regression and Missing Data

We use meta-regression to examine models of effect size heterogeneity

We assume here that we observe all of our effect sizes and their associated standard errors

Our missing data occurs on the moderators we want to use in our effect size model, moderators related to study characteristics such as participants, treatments, measures, etc.

To begin to answer the question of the appropriate analysis for missing data in meta-analysis, we need to understand the mechanisms for missing data in meta-analysis

Missing Completely at Random and Meta-Analysis Data

Imagine that one of our moderators in a meta-regression is duration of the treatment

To assume that duration of treatment is MCAR, we need to assume that primary studies fail to report duration of treatment completely at random - that the value of this moderator is unrelated to missingness

Do we really believe that?

Missing at Random and Meta-Analysis

Missing at Random would imply that the reasons treatment duration is missing is related to another completely observed variable in the meta-analysis data

This assumption is weaker than MCAR - but still is problematic

How would we know if treatment duration is MAR?

Missing Not at Random and Meta-Analysis

Missing Not at Random assumes that duration of treatment is missing because of it value in the primary study, e.g., we are missing values for duration of treatment more often when the treatment was short

But again, how do we even know this?

Exploring missing data in a meta-analysis

Before we can even make a decision about the appropriate way to handle missing data in a meta-analysis, we need to understand the nature of the missing data in our meta-analysis

In this presentation, we present strategies to explore the missing data in a meta-analysis data set, including looking at patterns of missingness among our moderators and effect size data

The rest of the presentation will discuss these exploratory techniques and preview ongoing work on missing data and meta-analysis

02

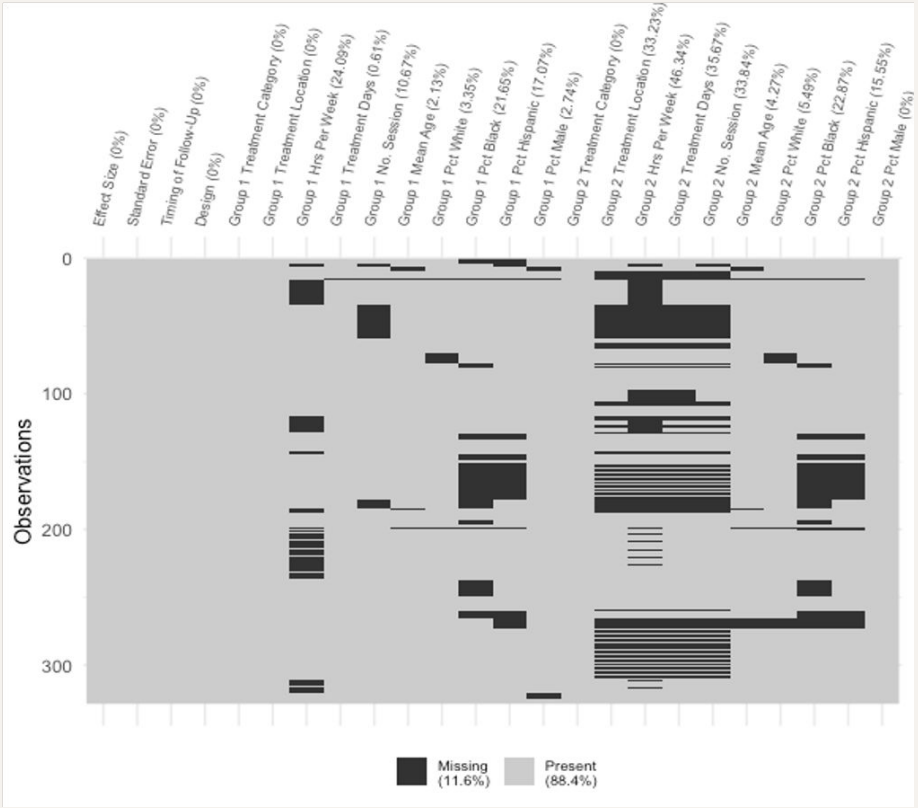
Understand Missingness

Exploratory Missingness Analysis

Questions we could explore with EMA

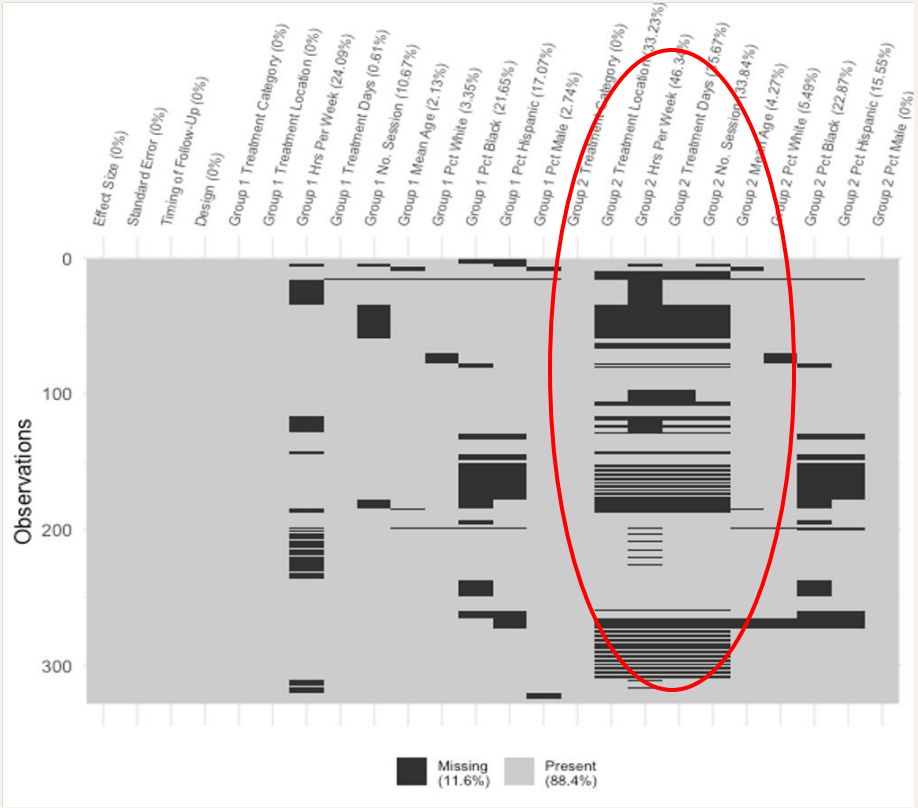
- Where information is scarce. How much data is missing in each covariate?
- What should we think about **bias of omitting cases**?
- Is missingness in one variable related to observed values in another variable?
- What **assumptions about missingness mechanisms** (MAR, MNAR, MCAR) are feasible in this case?
- Can we use meta-analytic methods, such as complete cases, which assume that data are MCAR?

Exploratory missingness analyses



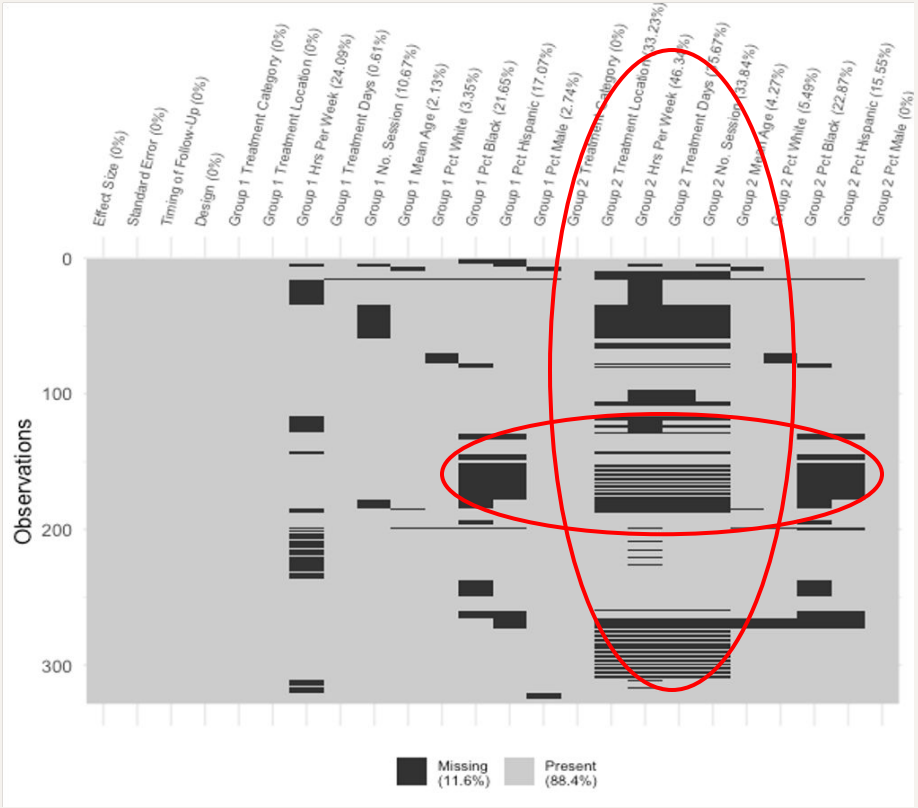
- Tanner-Smith et al. (2016) Study. Examined the impacts of substance abuse interventions for adolescents on subsequent substance use.
- The effect size used in the meta-analysis compares two groups of study participants (Group 1 & Group 2).
- Authors encounter missingness among their covariates.

What is this plot telling us about missing data?



- Explore the severity of missingness in the data (11.6%).
- Group 2 shows greater missingness.

What is this plot telling us about missing data?



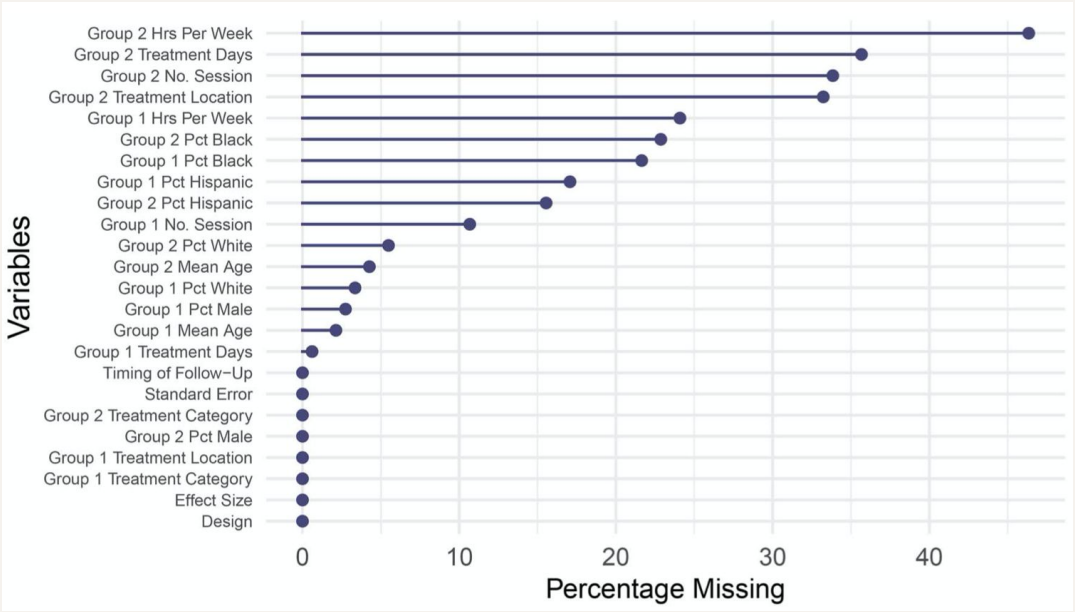
- Explore the severity of missingness in the data (11.6%).
- Group 2 shows greater missingness.
- Cases where covariates are missing for both groups (Treatment Intensity & Demographic information).

Amount of missingness

- Identify variables that are driving missing data problems.

- 10 variables with at least **10% of missing cases**.

- **Precision-weighted percentage**. Effects that are missing any covariate make up 74% of the total precision in the data.



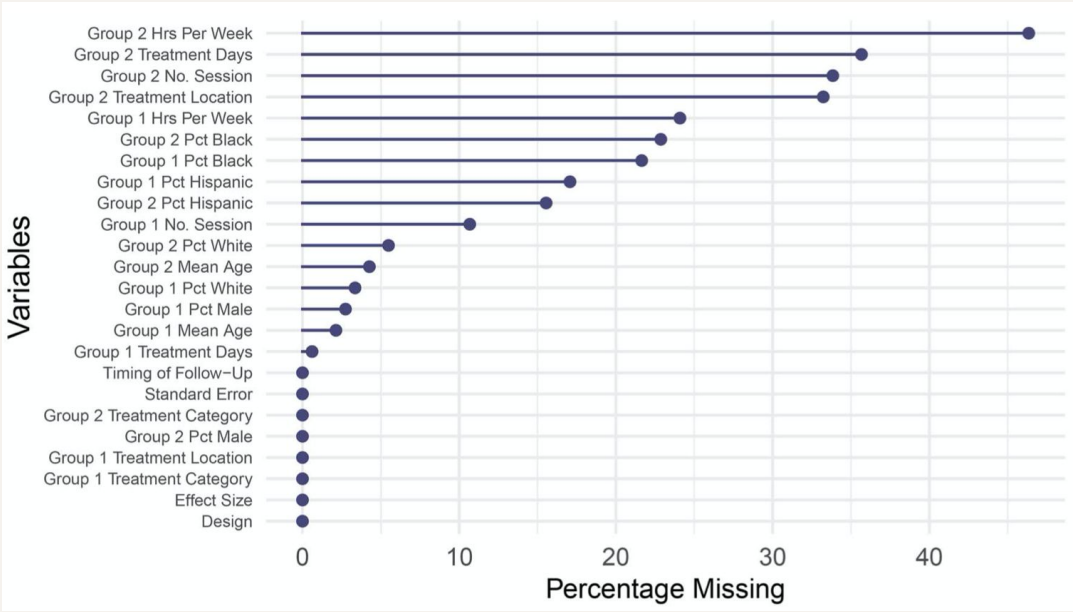
Amount of missingness

- Identify variables that are driving missing data problems.

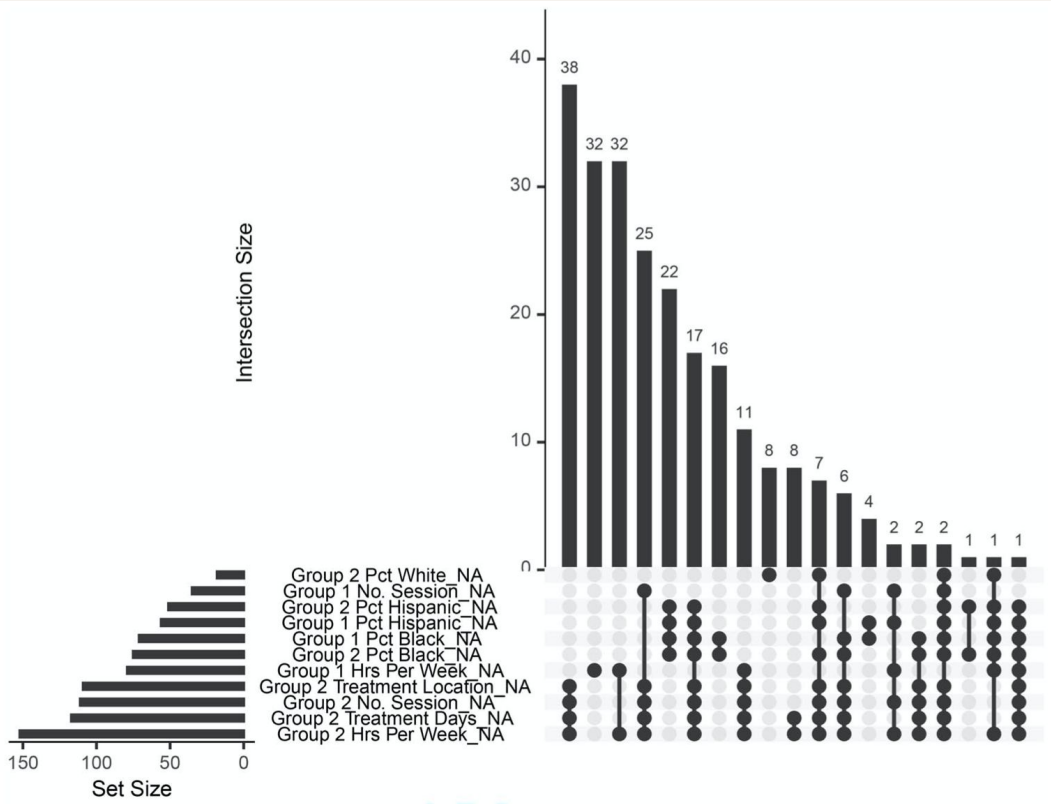
- 10 variables with at least **10% of missing cases**.

- **Precision-weighted percentage**. Effects that are missing any covariate make up 74% of the total precision in the data.

Relevant information when it comes to deciding which variable to include in the analysis.



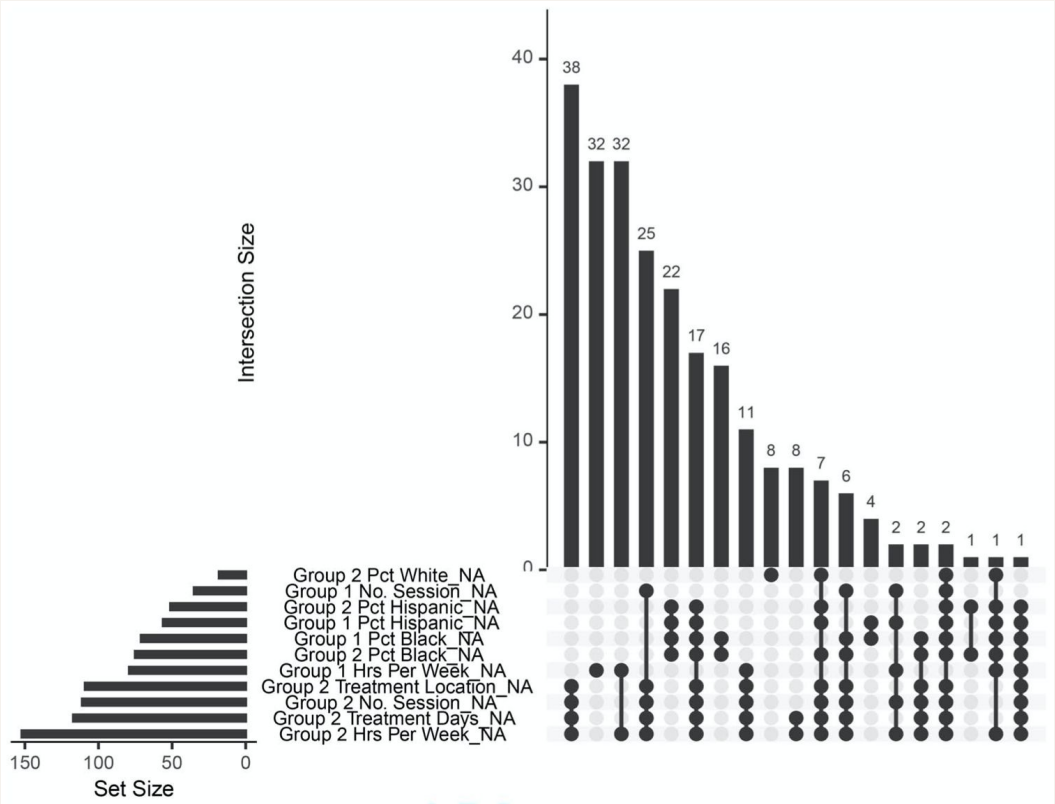
Identify Missingness Patterns



- Patterns where much of the information about the treatment condition in Group 2 is missing.

- Patterns that involve rows that are missing information about Groups 2's treatment as well as some demographic information.

Identify Missingness Patterns



- Patterns where much of the information about the treatment condition in Group 2 is missing.

- Patterns that involve rows that are missing information about Groups 2's treatment as well as some demographic information.

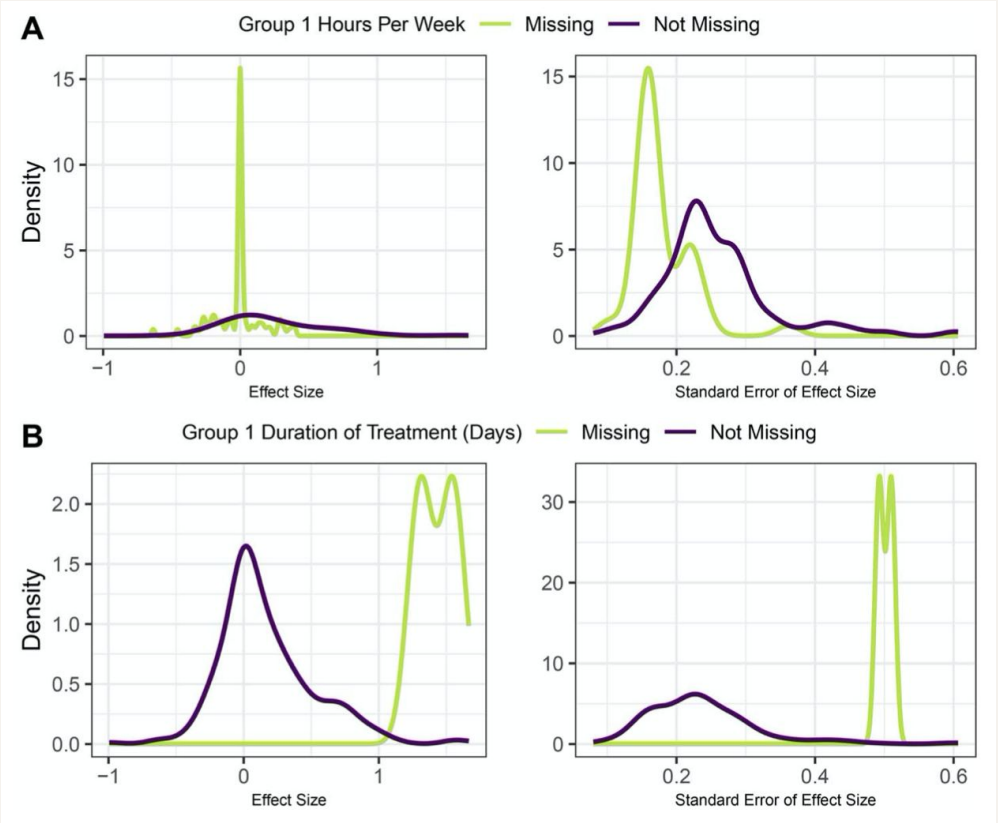
1. **Missingness in one variable is related to missingness in other missing covariates. Potential issue for MI.**

2. **Using Complete Case Analysis could lead to a large reduction in sample size.**

Is missingness correlated to observed variables?

- Figure A: Effect estimates are smaller and have smaller SE when covariate is missing.

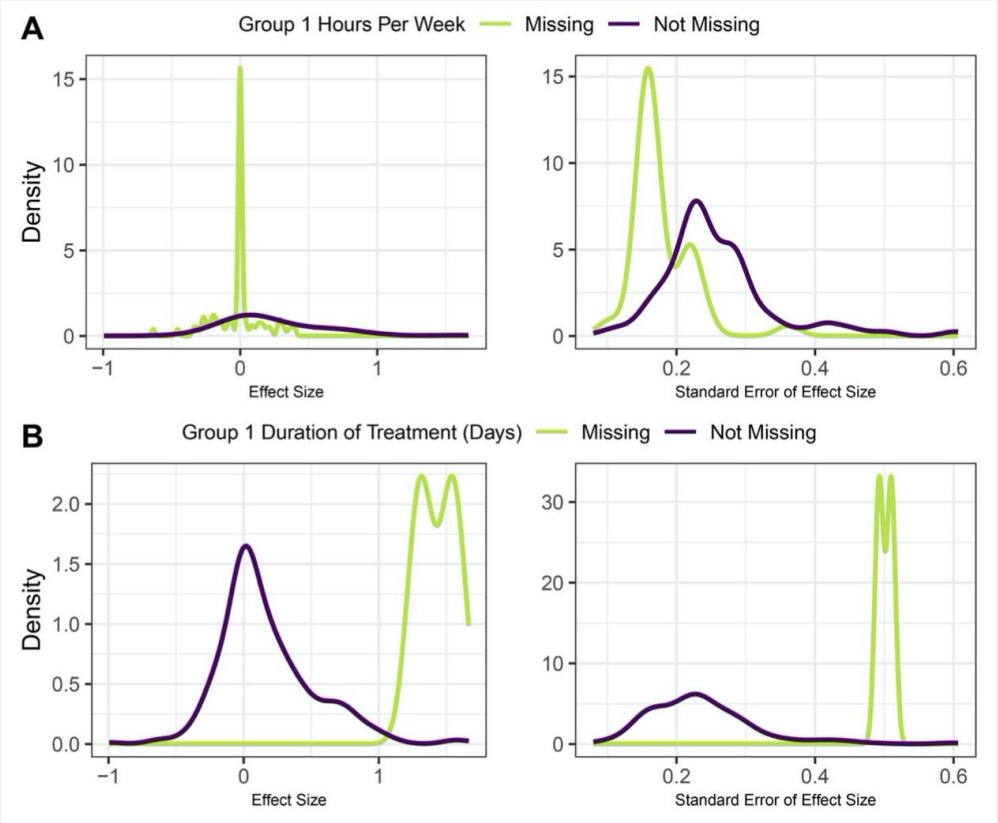
- Figure B: Effect estimates are larger and have larger SE when covariate is missing.



Is missingness correlated to observed variables?

- Figure A: Effect estimates are smaller and have smaller SE when covariate is missing.
- Figure B: Effect estimates are larger and have larger SE when covariate is missing.

1. **Omitting Effect Sizes with missing Group 1 Hours Per Week could lower the accuracy of the meta-regression.**
2. **MCAR is not a feasible assumption.**



Additional resources

- Schauer, J. M., Diaz, K., Pigott, T. D., & Lee, J. (2021). Exploratory Analyses for Missing Data in Meta-Analyses and Meta-Regression: A Tutorial. *Alcohol and alcoholism (Oxford, Oxfordshire)*, agaa144. Advance online publication. https://doi.org/10.1093/alcalc/agaa144

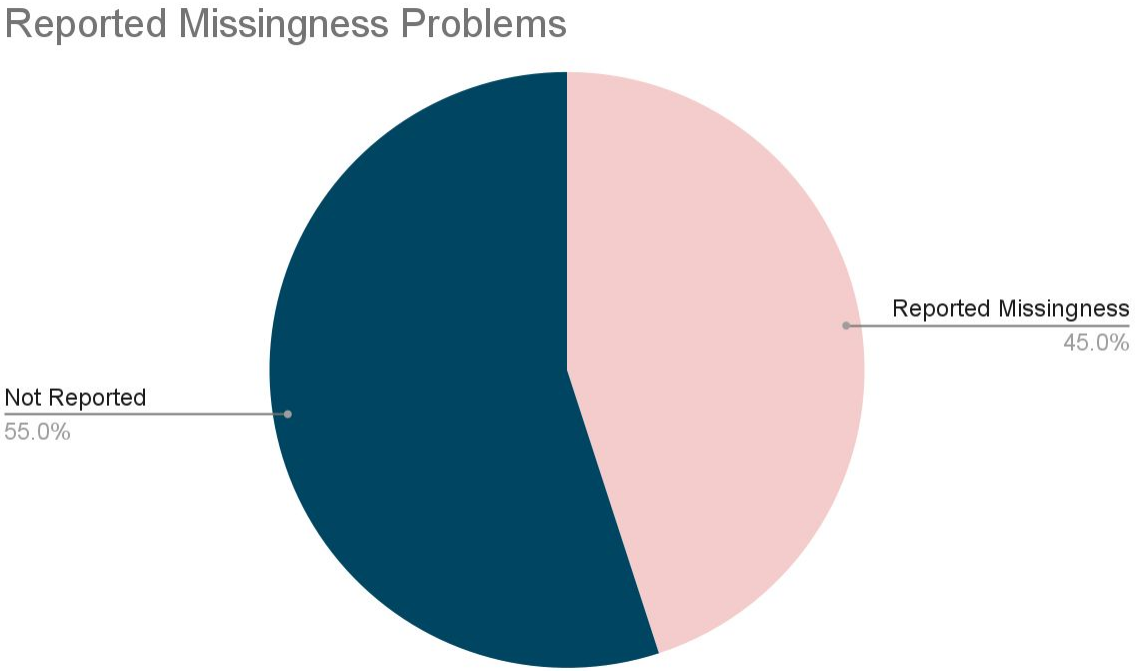
- Vignette – **bit.ly/2Z3pMM9**

03

Understand Current Practice

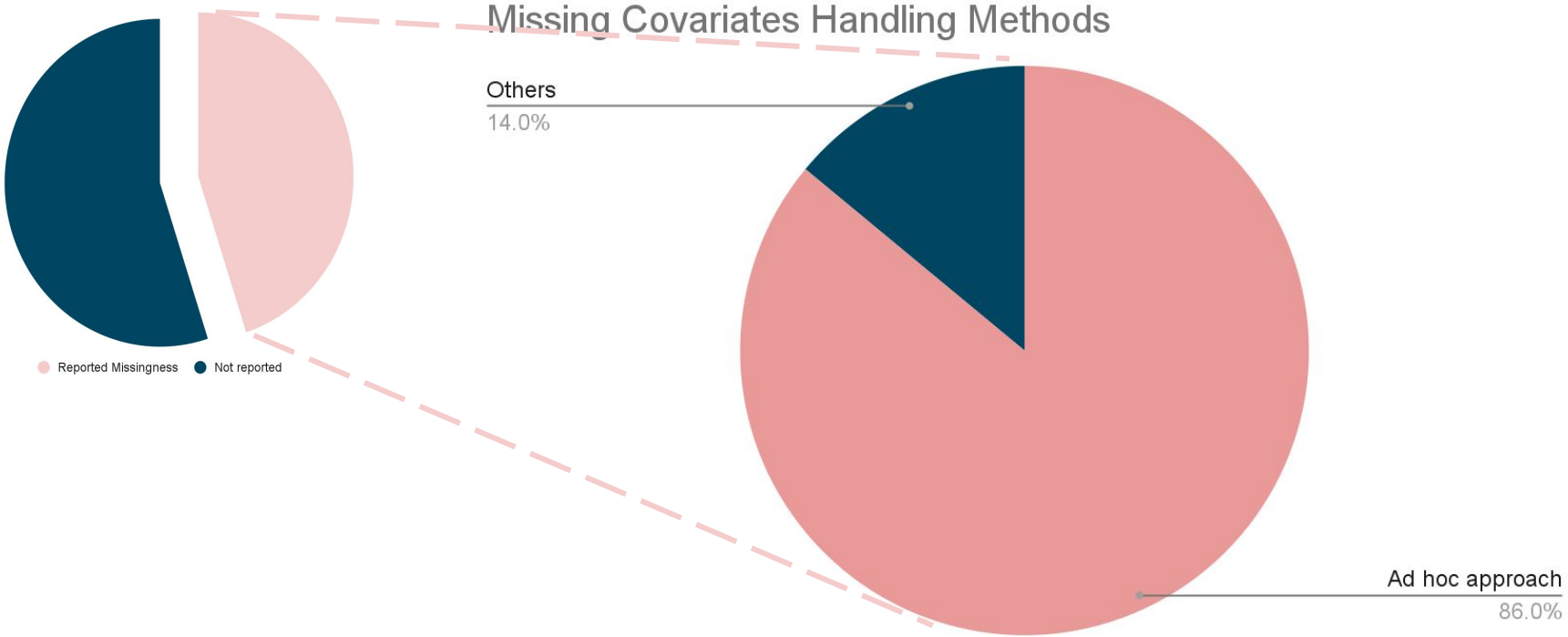
Complete-Case Analysis
Shifting-Case Analysis

Missingness problem is common in meta-analysis



Tipton, Pustejovsky, & Ahmadi (2019)

Current practice: Mostly ad hoc approach



Tipton, Pustejovsky, & Ahmadi (2019)

Current Practice for Handling Missing Covariates

Listwise Deletion

Pairwise Deletion
Shifting-units-of-analysis

- Complete-Case Analysis (CCA) & Shifting-Case Analysis (SCA)
- The impact of CCA and SCA has not been widely studied in meta-analysis literature.
- We have studied the conditions for the unbiased or biased estimates with CCA and SCA & have investigated the magnitude and sources of bias in meta-regression.
- This presentation focuses on the conceptual discussions.

CCA

Example: Single covariate meta-regression model

$$T_i = \beta_0 + \beta_1 X_i + u_i + e_i$$

The bias depends on:

(1) Missingness rates
(2) The total variations around the effect size estimate
(3) The relationship between *T* (effect size estimate) and the missingness in the covariate

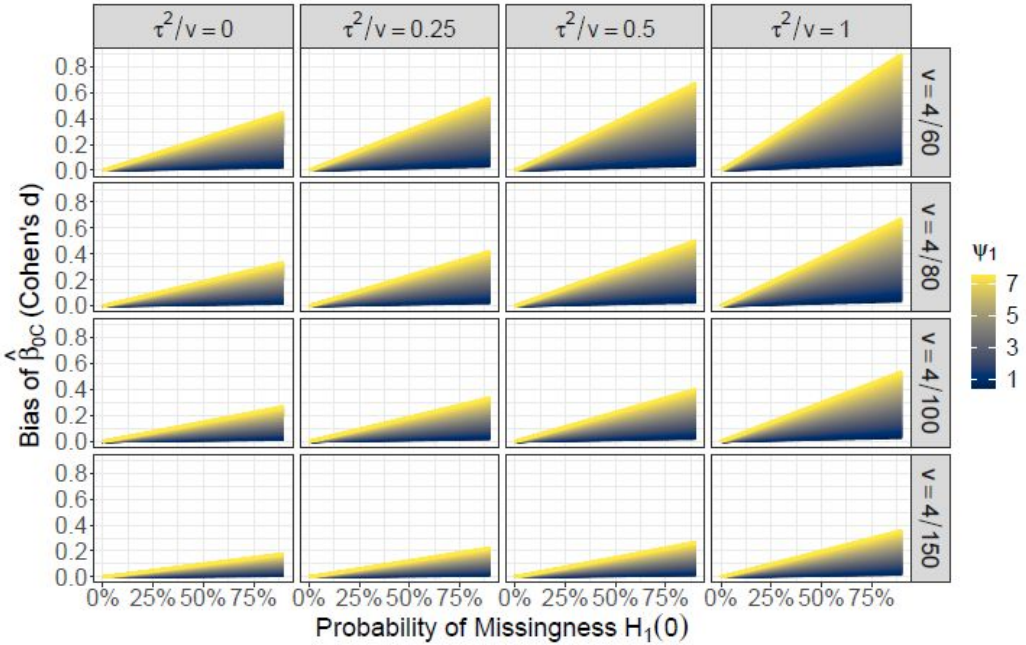
Does missingness depend on *T*?

(3) The relationship between *T* (effect size estimate) and the missingness in the covariate

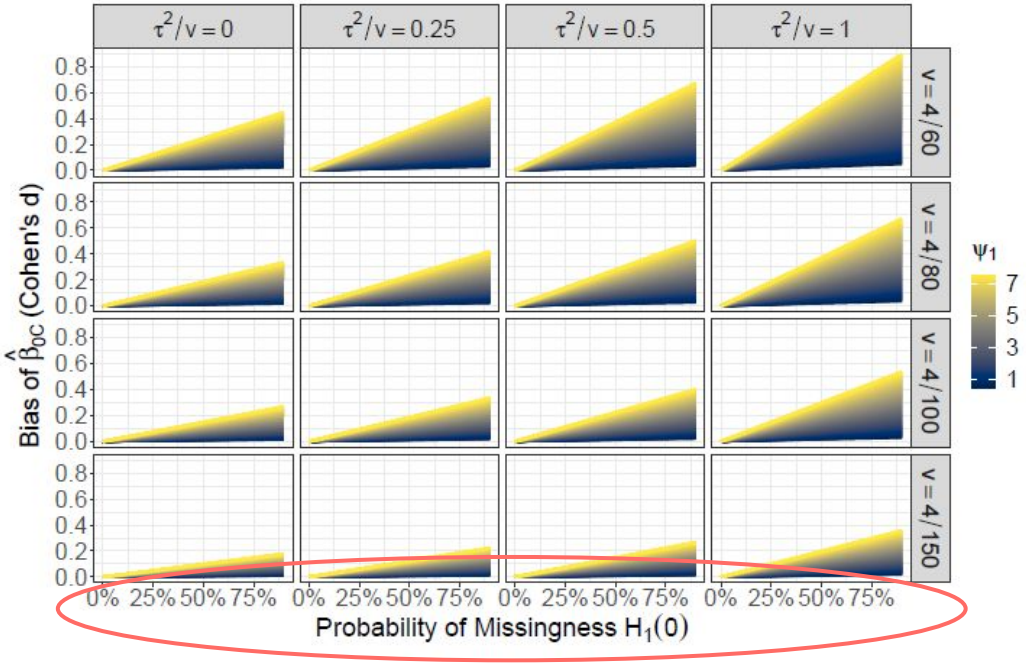
Selection model: $\text{logit}\left[p(\textit{Missingness in } X)\right] = \psi_0 + \psi_1 T + \psi_2 X + \ldots$

- Biased when the missingness is related to the effect sizes
  - Large bias with the strong relationship ψ_1
- The role of *T* in the selection model plays an important role, rather than the broader classifications of missingness mechanism (e.g., MCAR, MAR, MNAR).

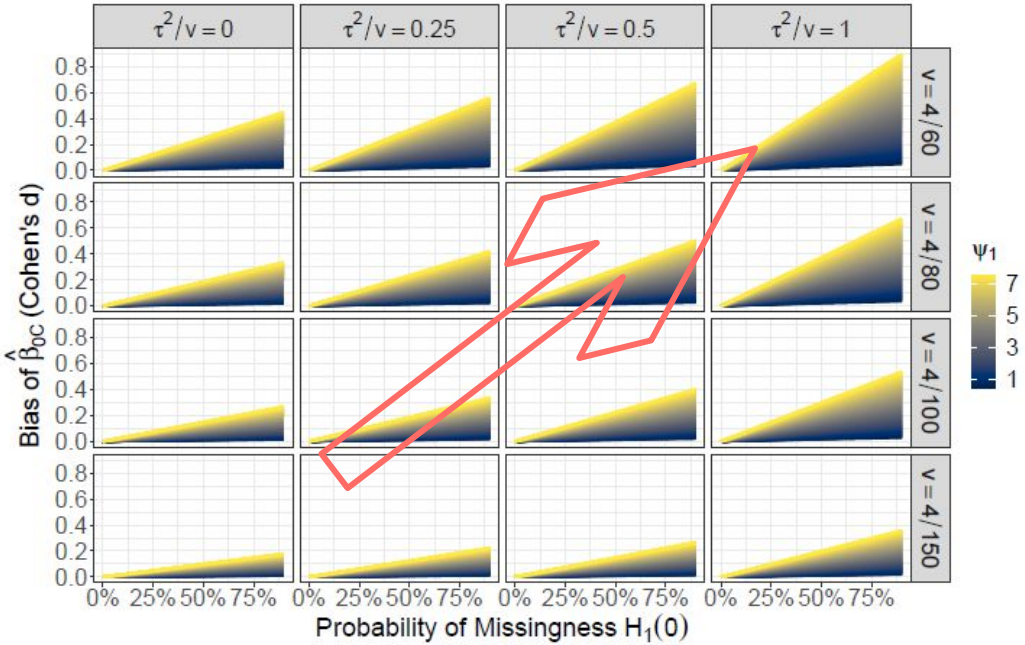
Bias depends on three factors:



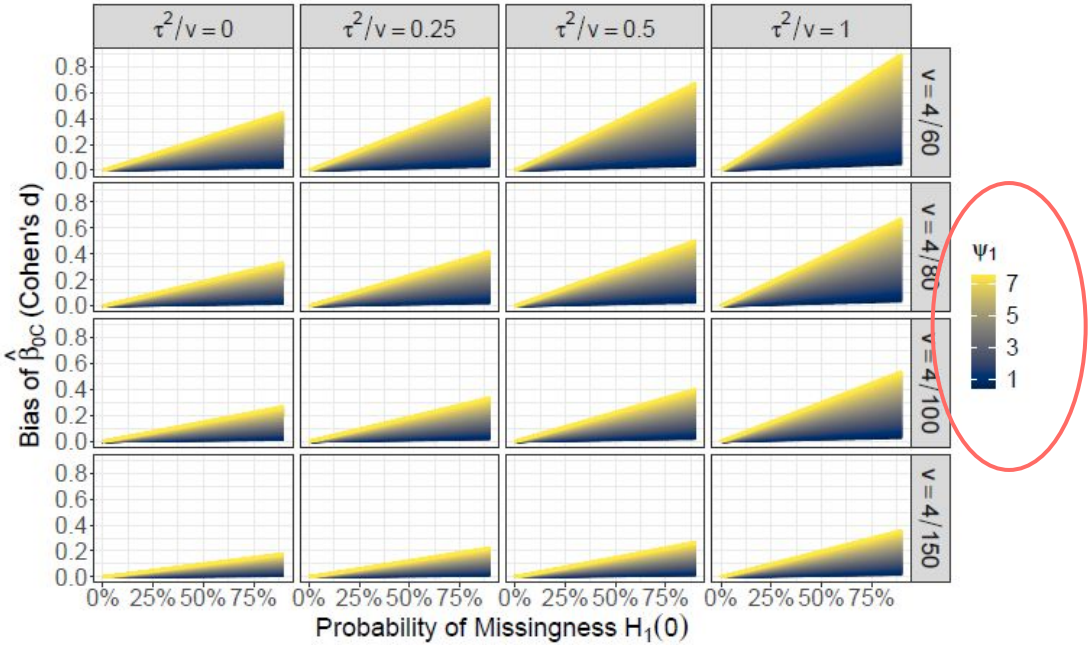
Bias depends on three factors: (1) Missingness rates



Bias depends on three factors: (2) Total variation



Bias depends on three factors: (3) ψ_1



SCA

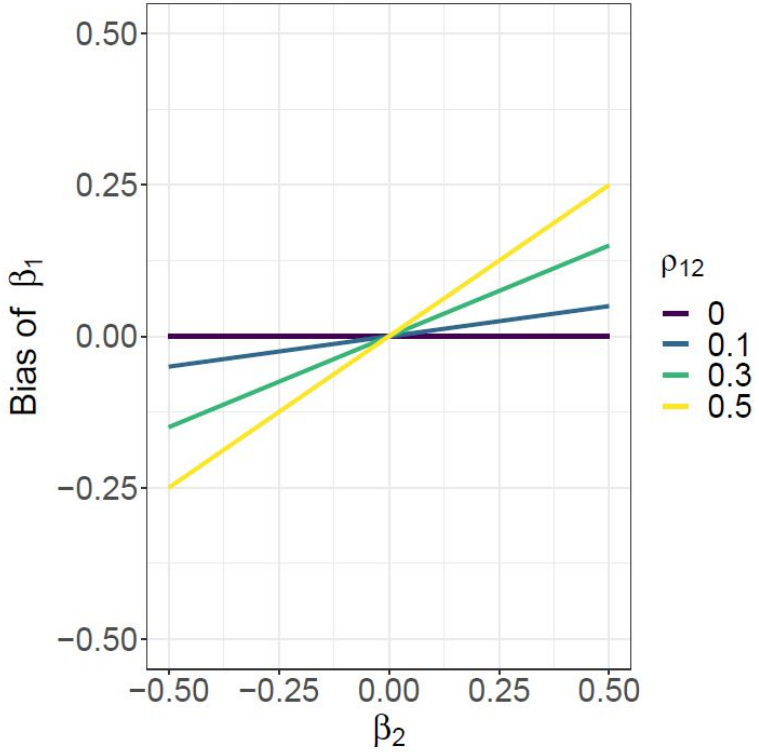
Example: Two-covariate meta-regression model

$$T_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i + e_i$$

The SCA fits two meta-regression models, one with X1 and another with X2.
Let's consider the first meta-regression model that only includes X1.

1. Two sources of bias in the SCA = The omitted variable bias (OVB) & missingness bias.
2. The OVB depends on:
   a. β_2 : the contribution of *X2* (omitted variable) to the complete-data model.
   b. *Cor(X1, X2)*: the correlation between the variable included and the variable omitted.
3. The missingness bias
4. Depending on the direction of the two sources of biases, SCA may or may not be more biased than CCA.

OVB in SCA


0.50
0.25
0.00
−0.25
−0.50
Bias of β_1
−0.50
−0.25
0.00
0.25
0.50
β_2
ρ_{12}
0
0.1
0.3
0.5

SCA

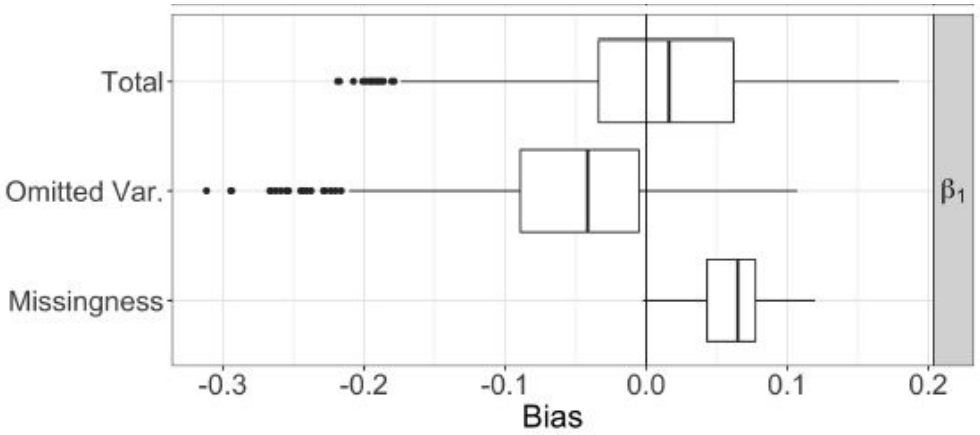
Example: Two-covariate meta-regression model

$$T_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i + e_i$$

The SCA fits two meta-regression models, one with X1 and another with X2.
Let's consider the first meta-regression model that only includes X1.

1. Two sources of bias in the SCA = The omitted variable bias (OVB) & missingness bias.
2. The OVB depends on:
   a. β_2 : the contribution of *X2* (omitted variable) to the complete-data model.
   b. *Cor(X1, X2)*: the correlation between the variable included and the variable omitted.
3. **The missingness bias = source of CCA bias**
4. **Depending on the direction of the two sources of biases, SCA may or may not be more biased than CCA.**

Bias in SCA = OVB + Missingness Bias



We need to ...

- Understand the missing data patterns via exploratory analysis: Relationship between effect size and missingness
- Be aware of potential bias in CCA and SCA
- Investigate the alternative missing data handling methods that have shown some promise for meta-regression. (Coming!)

04

Possible Solutions

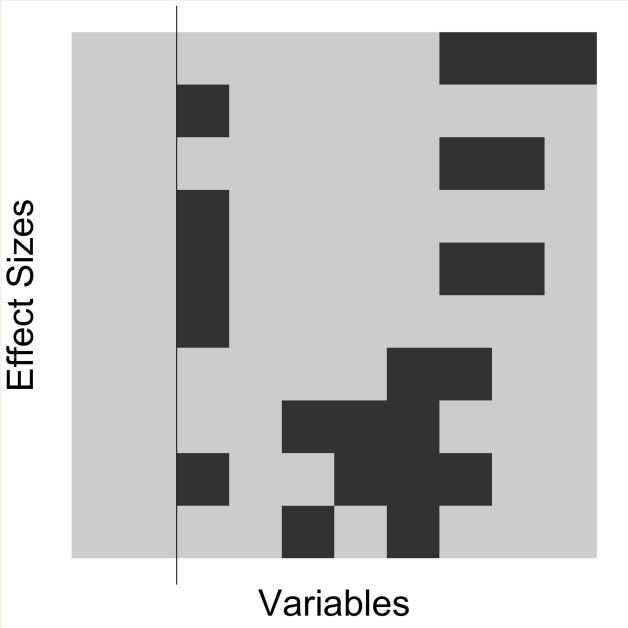
Several Possibilities

- Full information maximum likelihood (FIML)
  - SEM implementation: metaSEM (Cheung, 2021)
  - Assumes data are MAR

- E-M Algorithm (Dempster, Laird, & Rubin, 1977)
  - Assumes data are MAR

- **Multiple Imputation** (Rubin, 1987)
  - Flexible
  - Available in most programming languages (assuming MAR)
    - R: mice + metafor (Viechtbauer + VanBurren)

Multiple Imputation

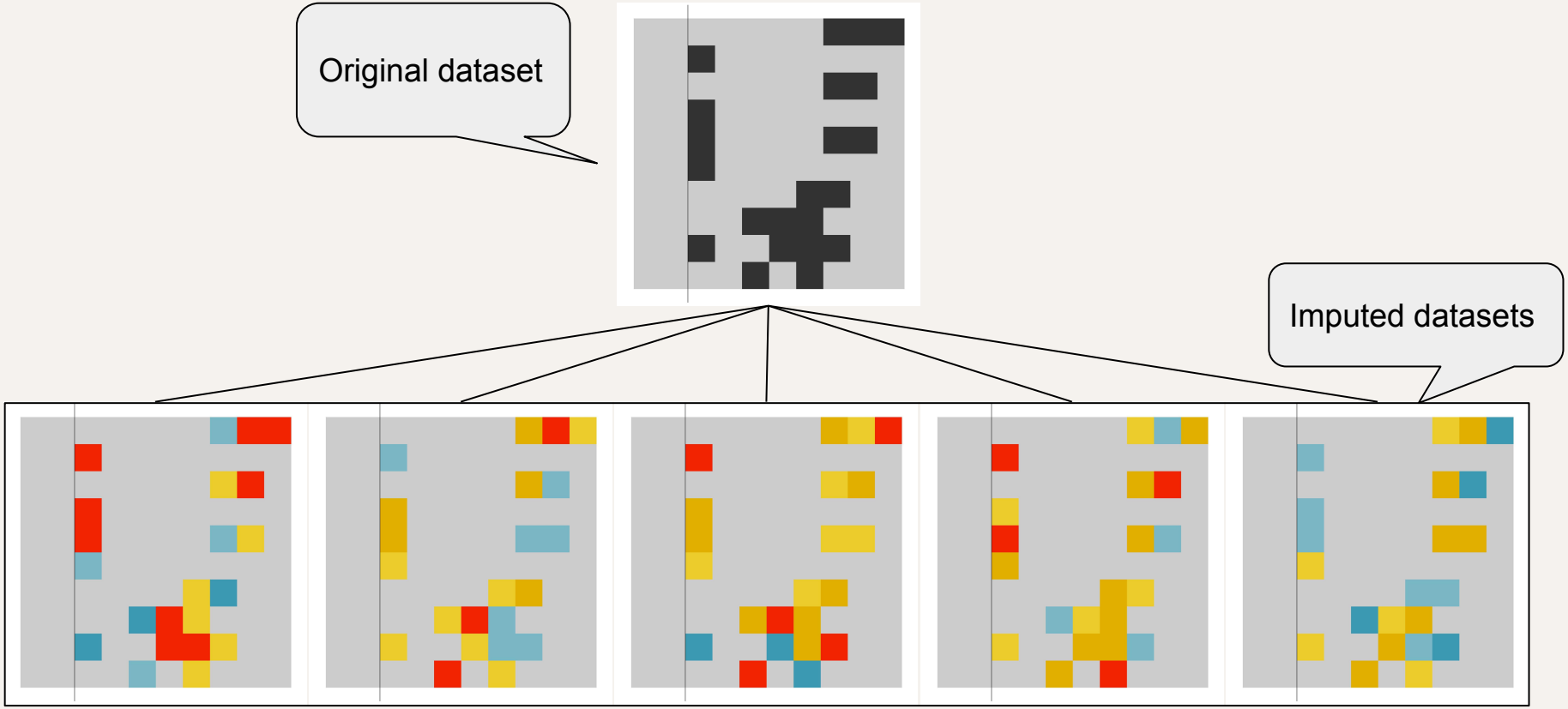


Fill in the missing values with what we *might* have observed.

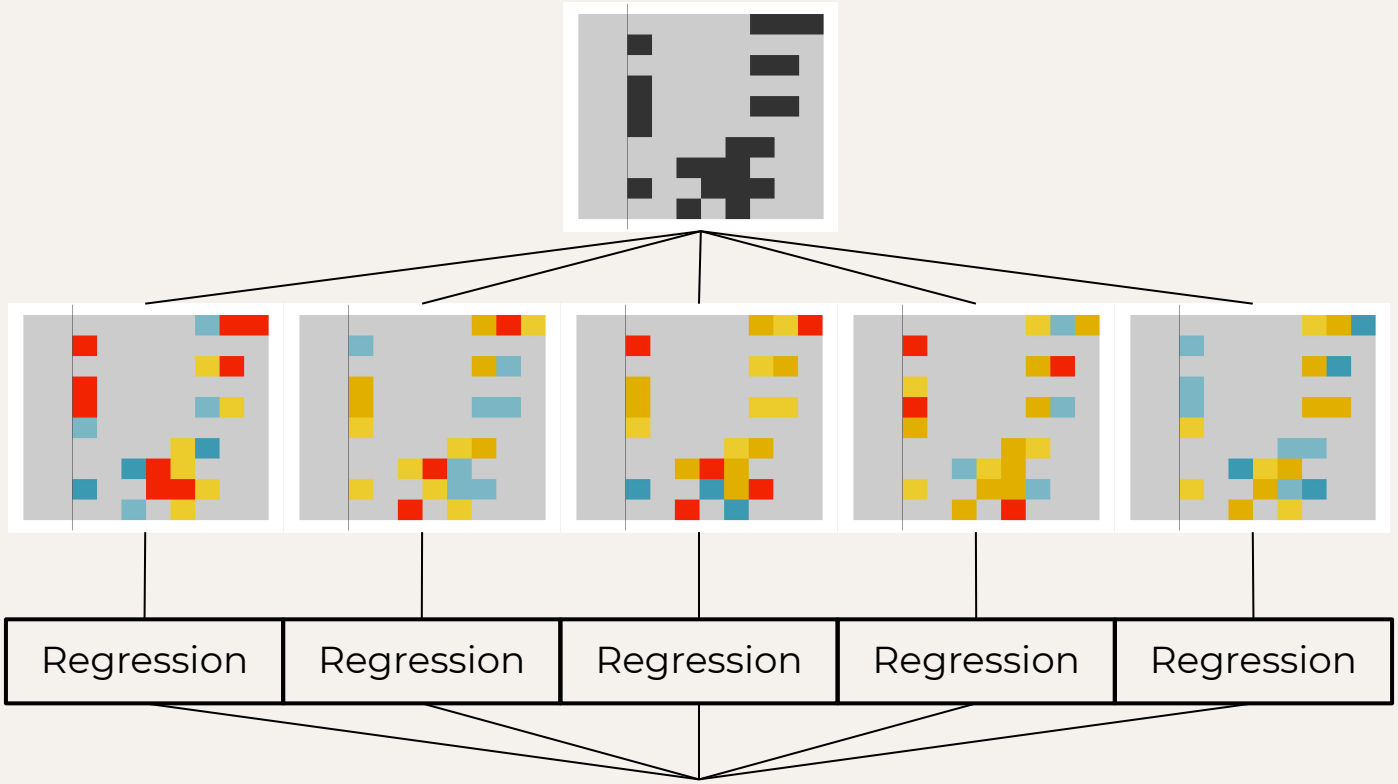
Fill in missing values multiple times since we don’t know what the missing values are.

Analyze data with different imputed values and pool results.

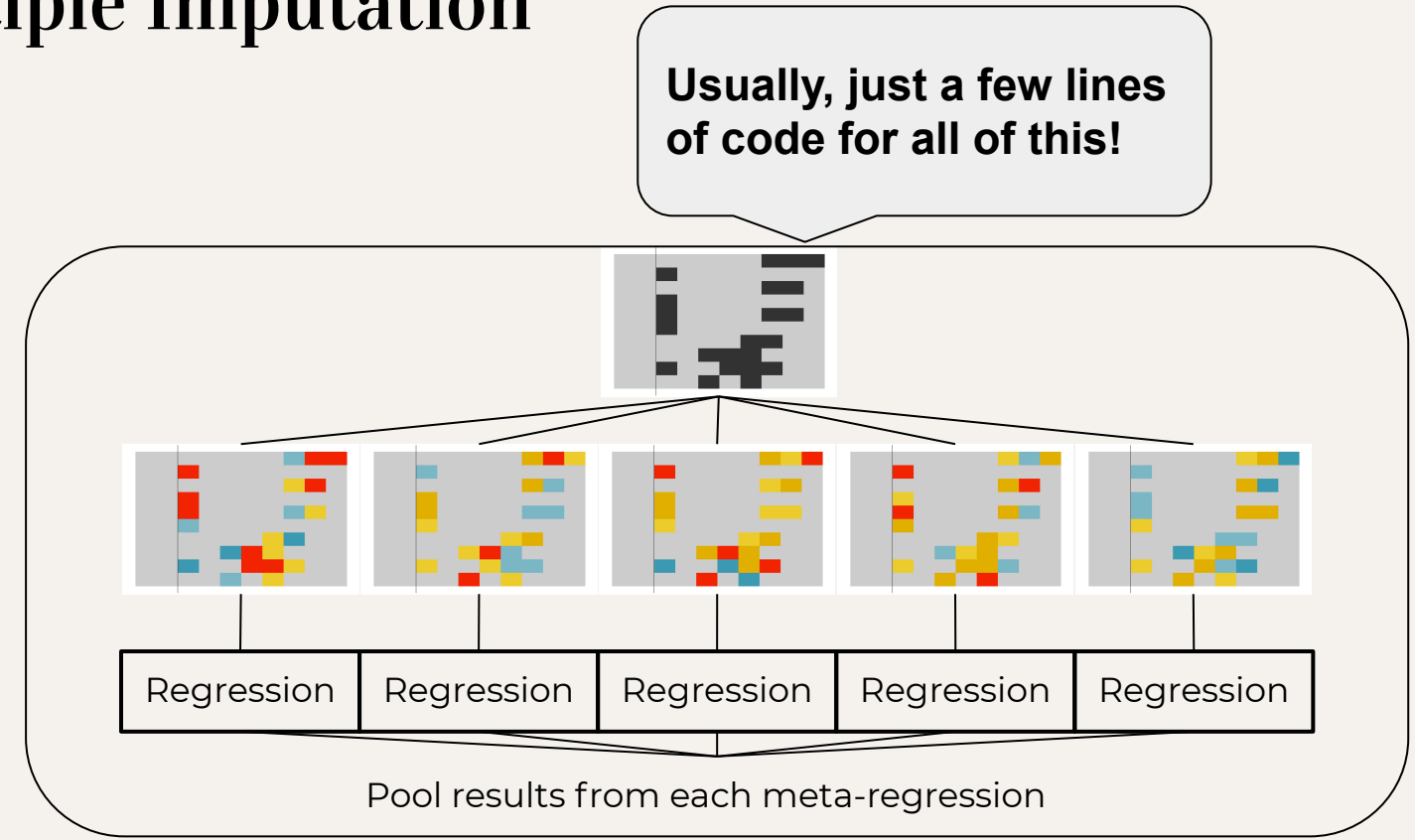
Multiple Imputation



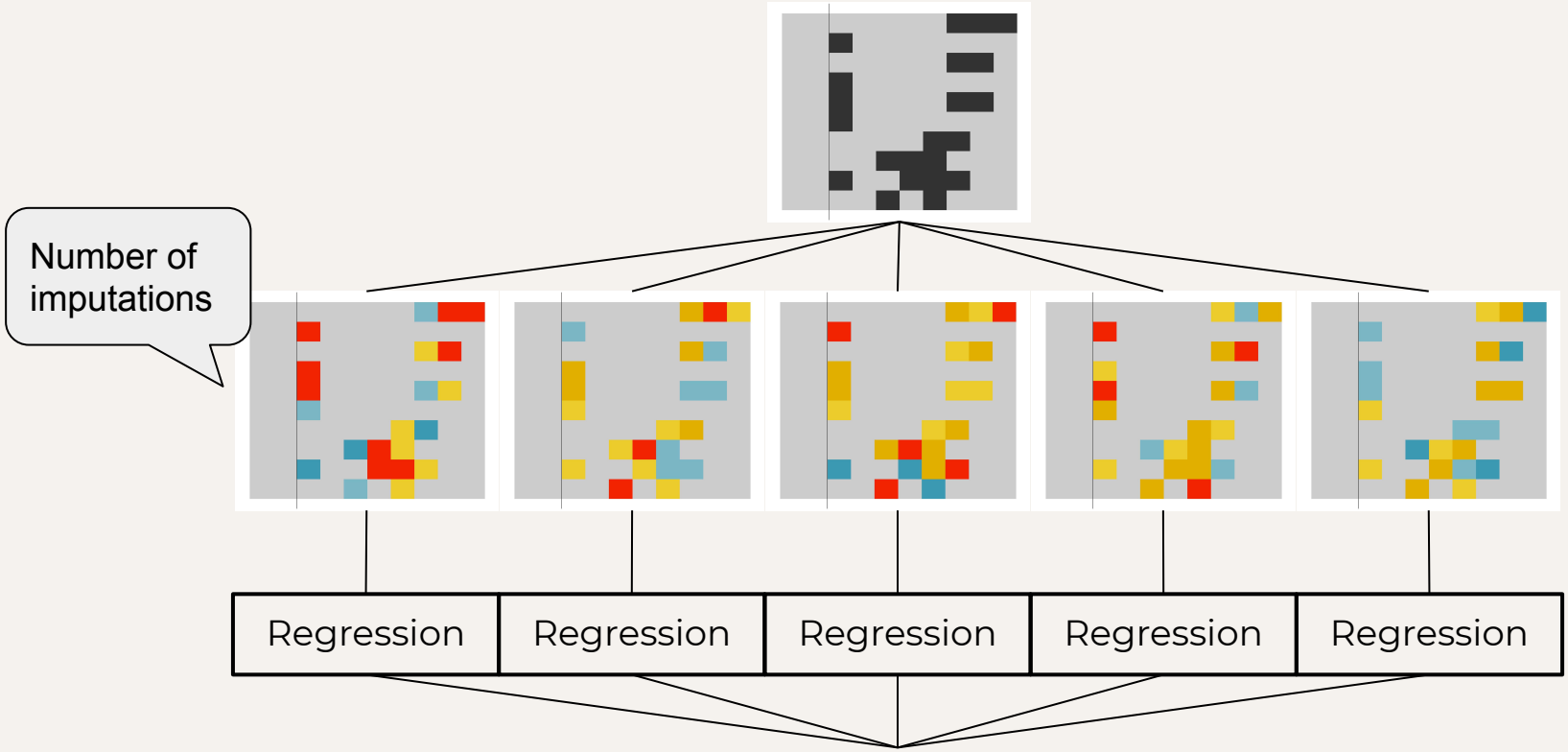
Multiple Imputation



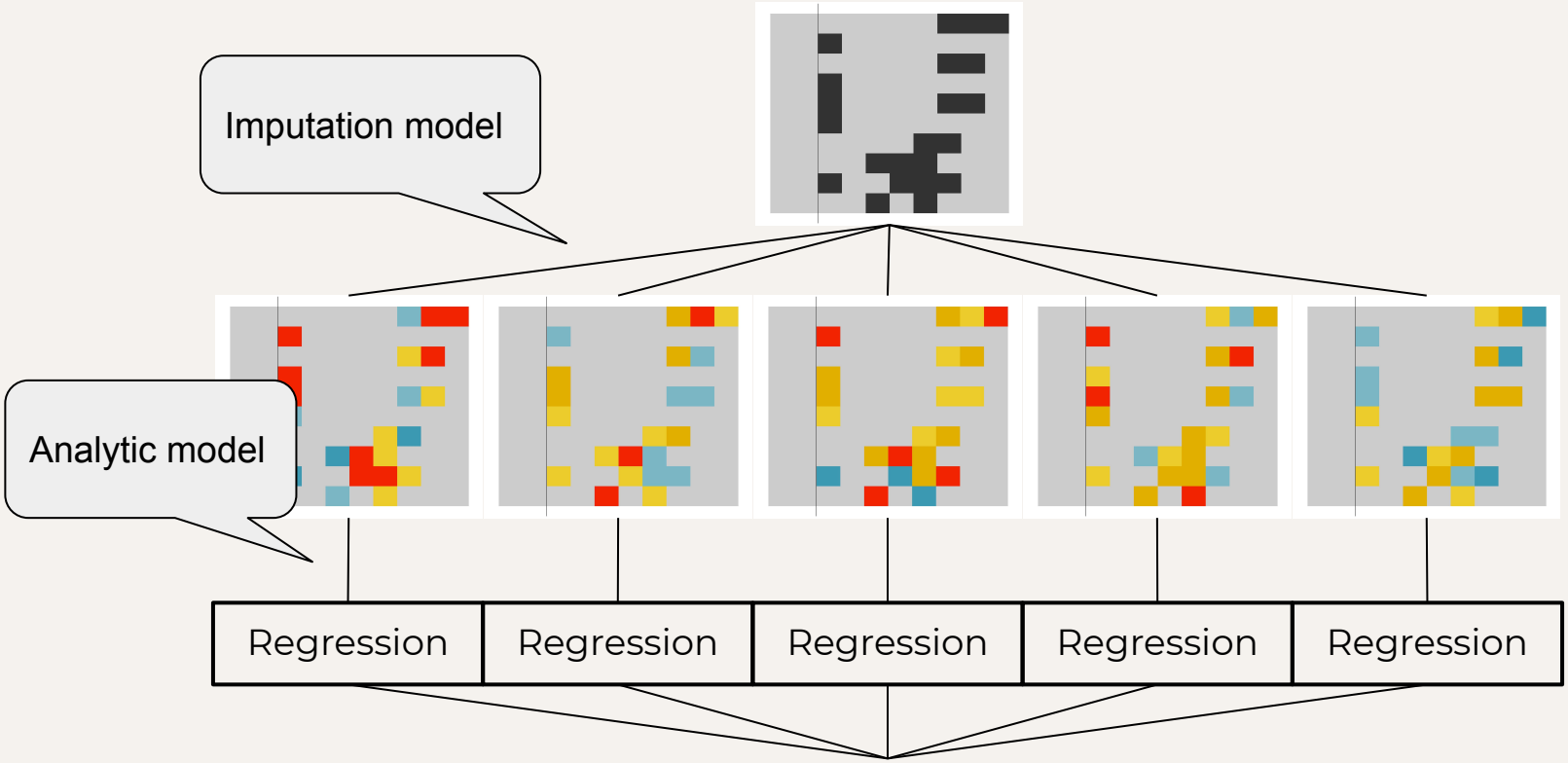
Multiple Imputation



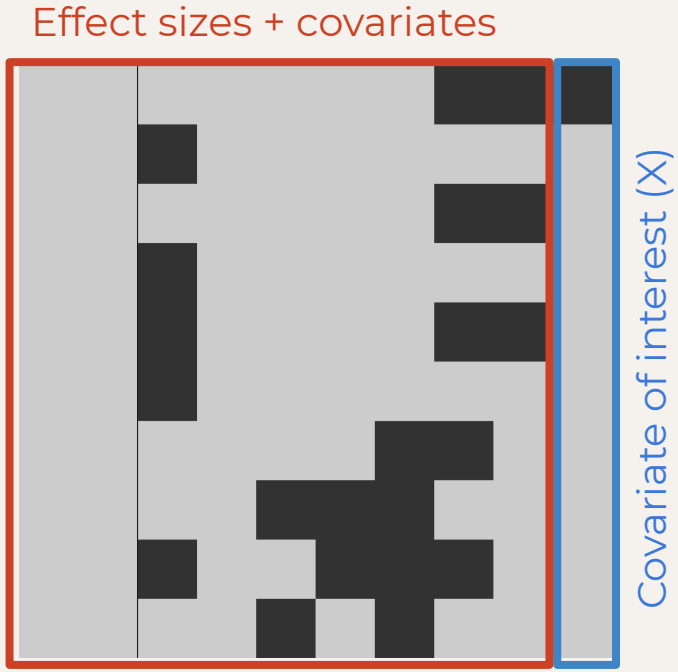
Number of Imputations



Two Model Types for Multiple Imputation



Imputation Models

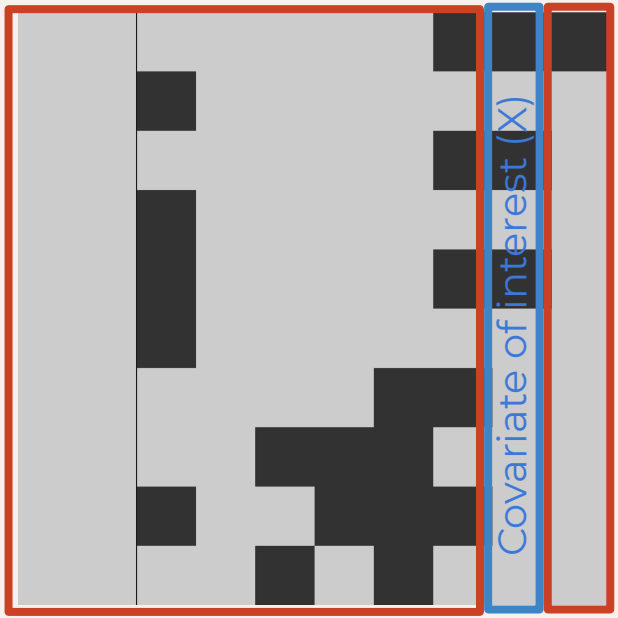


Regress X on effect sizes + SEs + other covariates

- Linear regression
- Logistic/probit regression
- CART
- Random forests
- GBM

Imputation Models

Effect sizes + covariates

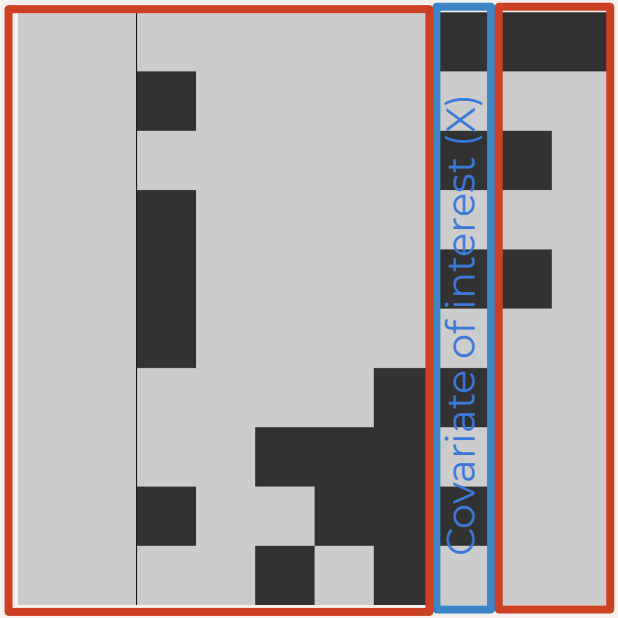


Regress X on effect sizes + SEs + other covariates

- Linear regression
- Logistic/probit regression
- CART
- Random forests
- GBM

Imputation Models

Effect sizes + covariates



Regress X on effect sizes + SEs + other covariates

- Linear regression
- Logistic/probit regression
- CART
- Random forests
- GBM

Modelling Considerations

MI is better at reducing bias and variance of regression estimates when the imputation model:

1. Is highly predictive of missing values
    a. Use as many relevant predictors in your imputation model as is feasible.
    b. Where possible, use flexible imputation models.

2. Accounts for known or theorized factors related to missingness
    a. Prioritize predictors you think are highly relevant.
    b. Potentially use effect size estimates as predictors.

3. **Is compatible with the analytic model**

Modelling Considerations

MI is better at reducing bias and variance of regres
imputation model:

Most MI software (e.g., R's mice) let you specify which predictors and models to use!

1. Is highly predictive of missing values
    a. Use as many relevant predictors in your imputation model as is feasible.
    b. Where possible, use flexible imputation models.

2. Accounts for known or theorized factors related to missingness
    a. Prioritize predictors you think are highly relevant.
    b. Potentially use effect size estimates as predictors.

3. **Is compatible with the analytic model**

Modelling Considerations

MI is better at reducing bias and variance of regression estimates when the imputation model:

1. Is highly predictive of missing values
    a. Use as many relevant predictors in your imputation model as is feasible.
    b. Where possible, use flexible imputation models.

2. Accounts for known or theorized factors related to missingness
    a. Prioritize predictors you think are highly relevant
    b. Potentially use effect size estimates

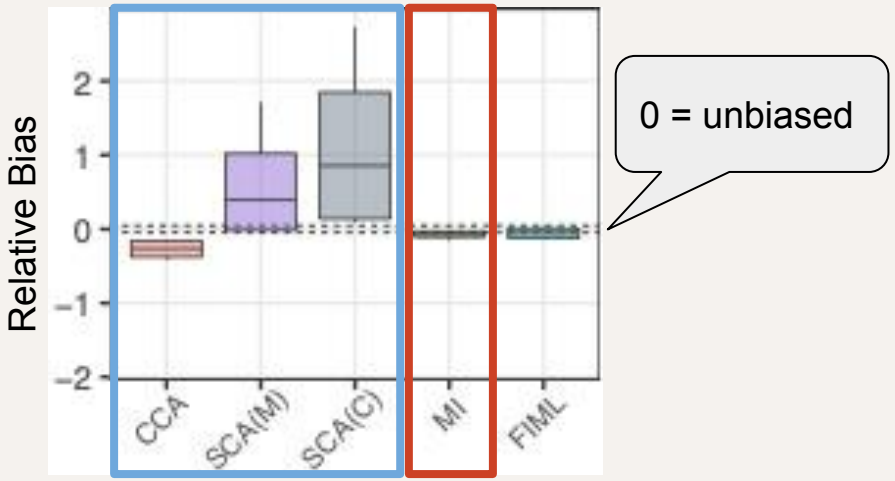
3. Is compatible with the analytic model

Coming soon!

MI without compatibility is still pretty good!

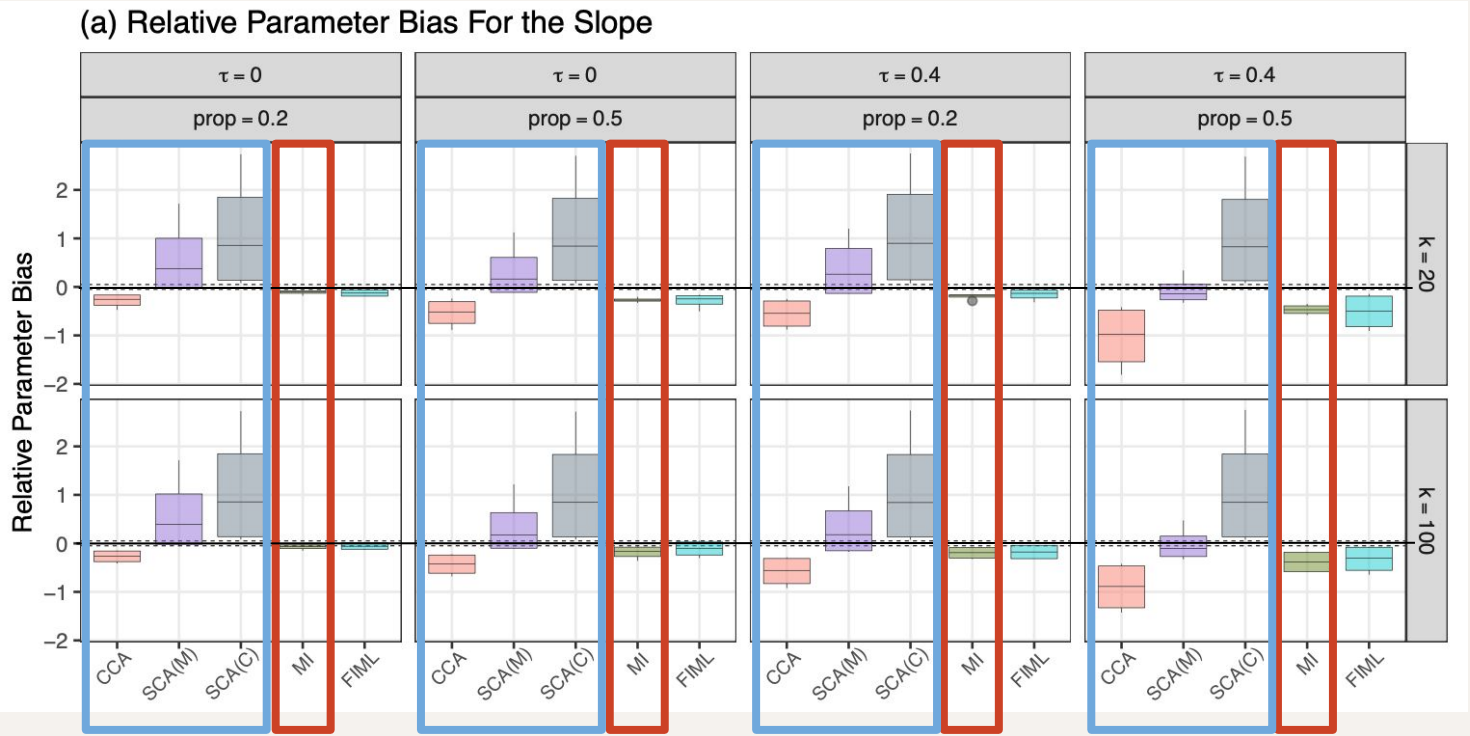
Simulations by Lee & Beretvas:

- *k* = 100 effect sizes
- 20% are missing a covariate
- Very little residual variation
- Analysis uses
  - **CCA**
  - **SCA**
  - **MI**
  - FIML



Lee & Beretvas, in progress

MI without compatibility is still pretty good!

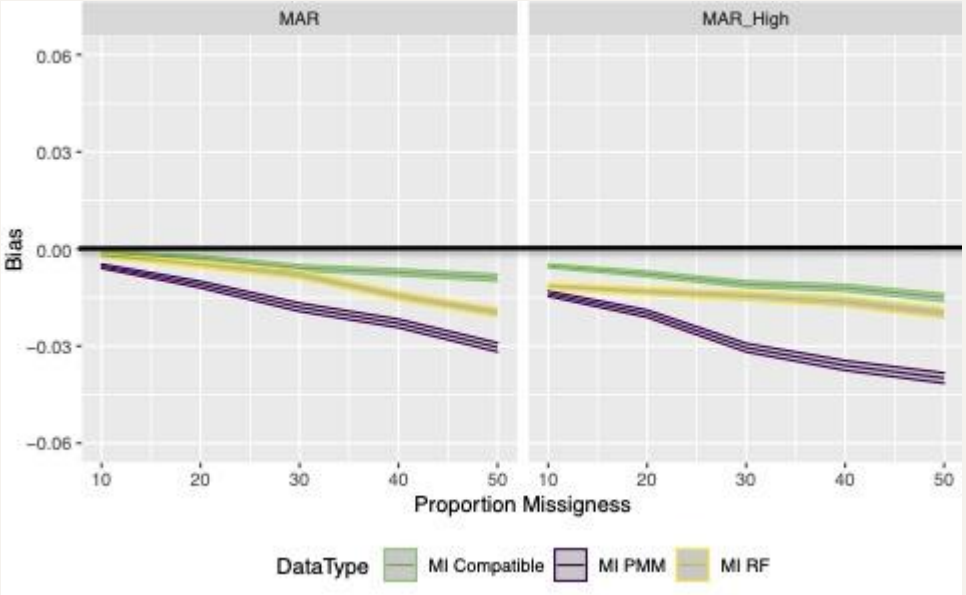


Lee & Beretvas, in progress

Compatible MI is even better!

Simulations by Diaz (2021)

- Strong vs. weak correlation with effect size
- Incompatible imputation
  - **Random forests (RF)**
  - **Nearest neighbor (PMM)**
- **Compatible imputation**
  - Reduces bias further, especially with large amount of missingness!



Concluding Thoughts

- EMA can be extremely useful for understanding information missing in the literature, as well as possible issues for analyses.
    - Check missingness rates (>5-10%) by variable.
    - Check amount of missing precision.
    - Check missingness patterns.
    - Check for relationship between missingness and effect size estimates!

Concluding Thoughts

- CCA/SCA will often be biased.
  - Bias will arise for both MAR and MNAR data
  - Bias increases when
    - Greater proportion of missingness
    - Greater within- and between-effect variance
    - Stronger correlation between effect estimates and missingness

Concluding Thoughts

- MI offers improvements over CCA/SCA.
  - Use variables that are predictive of covariates that are missing.
  - Even without compatible imputations, MI offers improved accuracy and precision over CCA/SCA.
- **EMA -> mice -> metafor (+clubSandwich, robumeta)**
  - Further refinements for compatible imputation are forthcoming from all four of us!

Thank you!