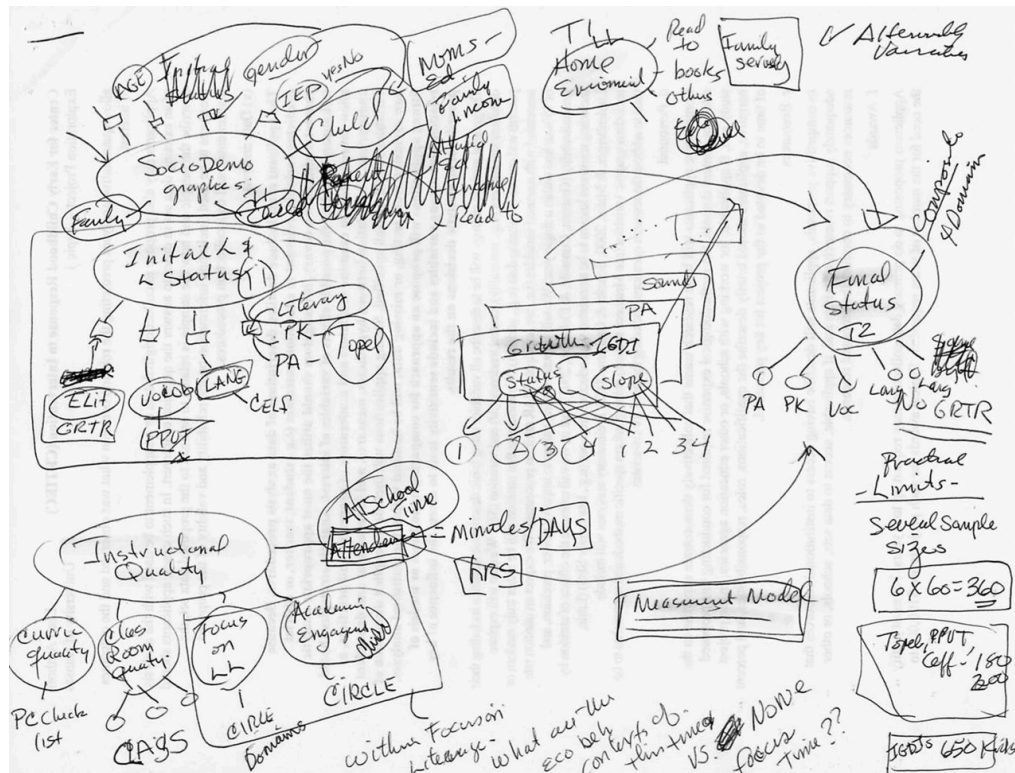# Handling Missing Data in Research: A Practical Guide

Introduction

Waylon Howard

Webinar, November 19, 2024

navigating research complexities

About me

Waylon Howard

- Principal Quantitative Methodologist and Biostatistician at *Biostatistics, Epidemiology and Analytics in Research (BEAR) Core* (Seattle Children's Research Informatics and Biostatistics)

  - Research methods, statistics, and measurement

  - Grant writing, manuscript development, and methodological innovation

  - Training and mentoring

About you

- What's your name?

- What is your research area?

- What are your experiences with missing data in research (and the tools used)?

- What are your expectations for this course?

Preliminaries

Slides and material are available at

https://github.com/wwwaylon/mi-2024

- The session consists of a combination of lectures and hands-on exercises

- Feel free to ask questions anytime

- We will primarily rely on `GitHub`, `R` & *RStudio* with additional support for `SAS` and even `Excel`.

Course schedule

Wednesday, November 20th, 2024

When?	What?
12:00 - 12:30	Introduction
12:30 - 1:00	Traditional Methods
1:00 - 1:45	Modern Methods
1:45 - 2:00	*Questions*

What is missing data?

Skip Questions   **Drop Out**   **Miss Some Assessments**   **Technical Issue**   **By Design**   **Human Error**

Basically, values for certain variables are unavailable.

Why missing data matters?

Missing data can significantly affect research outcomes, policy decisions, business strategies, and real-world interventions.

- **Bias** (systematic errors in statistical estimates)
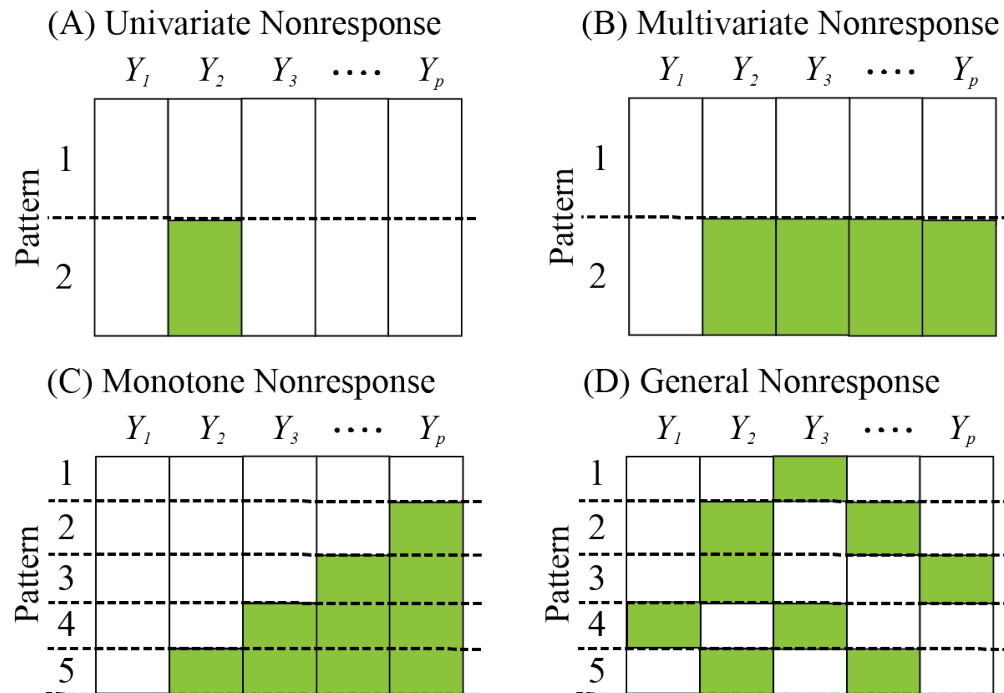- **Power** (decreased statistical power)

See Enders, C. K. (2022) for further details

Defining missing data

As with many concepts in science, missing data is defined and addressed in different ways. We will explore some basic definitions next.
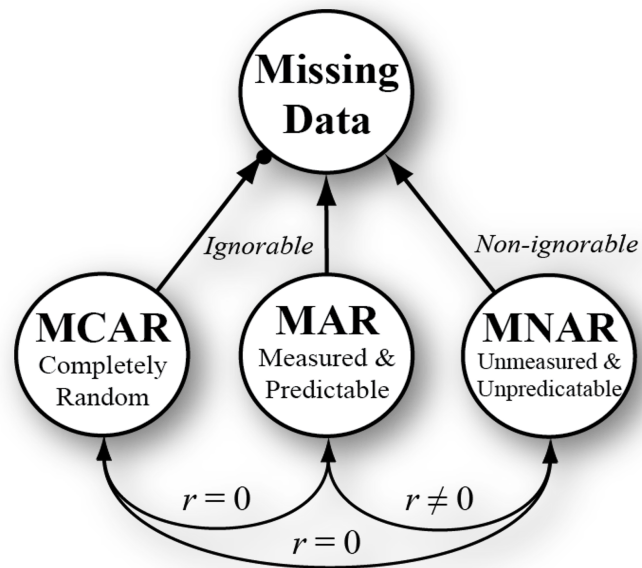
Defining patterns

Where are the data missing?

Four typical missing data patterns. Shaded areas represent missing data.

Defining mechanisms

Why are the data missing?

See *Rubin's Missing Data Classification System* for further details. Illustration adapted from Little, T. D. (2024).

MAR – MNAR continuum 😲

These three kinds of missingness should not be thought of as mutually exclusive categories of missingness, despite the fact that they are often misperceived as such... The best way to think of all missing data is as a continuum between MAR and MNAR (Graham, J. W., 2009, p. 567).

PRO dataset

Descriptive Statistics

- var1 ($M = 50.42$, $SD = 7.65$)

- var2 ($M = 46.25$, $SD = 6.52$)

- $r = 0.85$, $p < .001$

Data sources: xlsx, sas, R

ID	var1	var2
1	36.60	40.00
2	41.80	40.00
3	42.60	35.00
4	43.10	43.00
5	43.40	42.60
6	44.20	39.00
7	44.90	45.00
8	46.30	45.20
9	48.60	50.00
10	49.00	40.00
11	50.00	45.00
12	51.60	44.70
13	54.60	48.50
14	54.80	50.20
15	55.70	42.80
16	57.20	50.00
17	57.60	54.00
18	60.30	52.00
19	60.90	58.00
20	65.30	60.00

ID	var1	var2
1	36.6	40.0
2	41.8	40.0
3	42.6	35.0
4	43.1	43.0
5	43.4	42.6
6	44.2	39.0
7	44.9	45.0
8	46.3	45.2
9	48.6	50.0
10	49.0	40.0
11	50.0	45.0
12	51.6	44.7
13	54.6	48.5
14	54.8	50.2
15	55.7	42.8
16	57.2	50.0
17	57.6	54.0
18	60.3	52.0
19	60.9	58.0
20	65.3	60.0

Complete Data

$R = 0.85$, $p < 0.001$
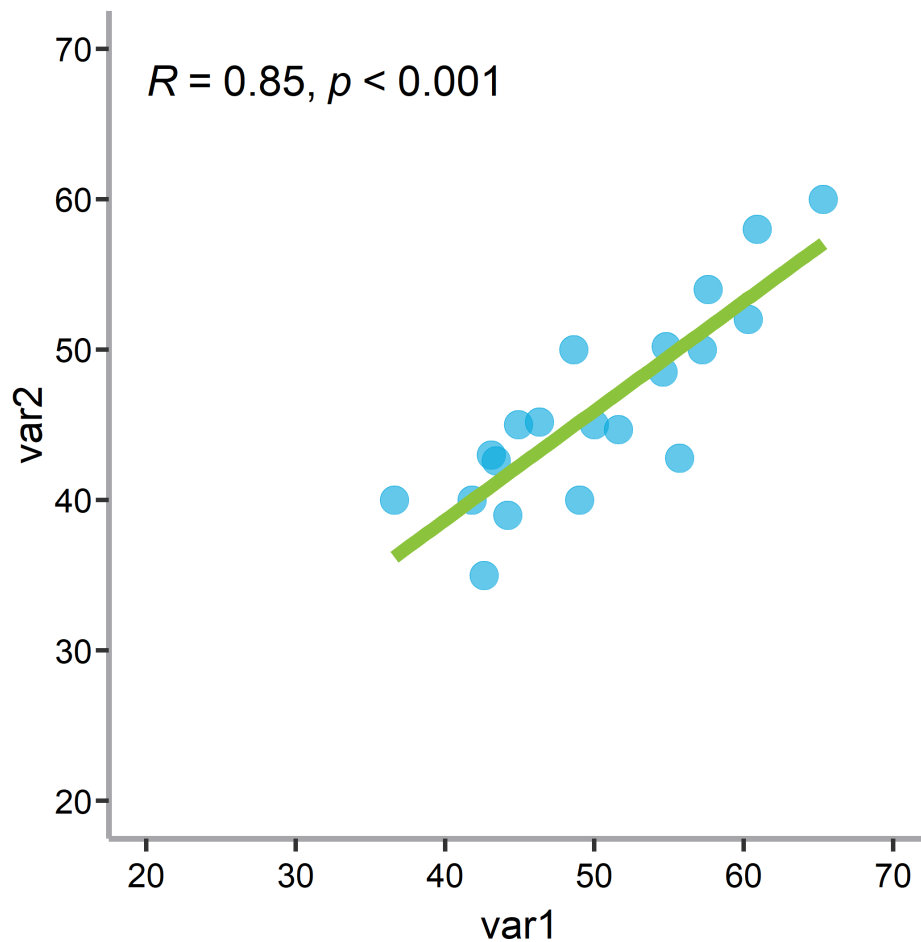
Traditional Methods

We will highlight some traditional ways of thinking about and approaching the problem of missing data. While these techniques are straightforward, they have important limitations to consider.

- listwise (or pairwise) deletion
- mean imputation
- last observation carried forward
- ...
- single imputation
- stochastic imputation

Modern Methods

We will then get into more sophisticated methods, such as Full Information Maximum Likelihood (FIML) and Multiple Imputation (MI) which offer more robust solutions by using all available data and accounting for uncertainty associated with missing values, leading to more reliable and valid inferences.[1]

[1] Although the algorithms used by stats programs can be quite complex, I will focus on presenting the core concepts in a way that's easy to understand.

Any questions so far?