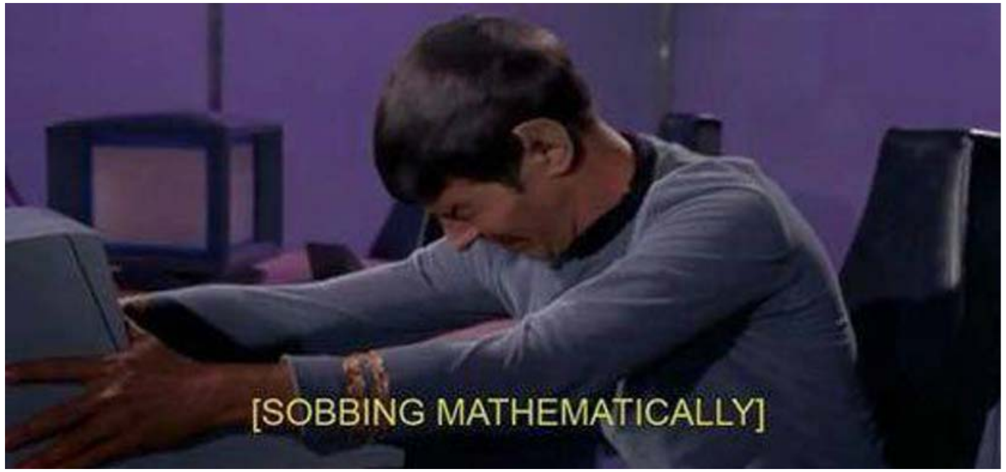
Missing Data Makes Me Sad

If we only use complete cases (i.e., listwise deletion):

1. Missing data cause a loss of efficiency and makes everyone sad
2. Results from the non-missing data may be biased and that's a waste of time and also sad



Takeaways:

- Understand how missing data can be problematic and how they can be addressed
- Identify the three common classifications of missing data (**MAR/MNAR/MCAR**) and how they differ
- Understand how multiple imputation (MI) is one robust way to deal with missing data and the general steps involved in this process
- What are the different sources of **uncertainty** related to MI
- How to examine proportion and patterns of missing data visually using plots
- What is convergence and how is it relevant to MI
- Understand how to analyse MI datasets and interpret their pooled results

MONASH University

Missing Data Assumptions

Missing data often are classified:

- **Missing completely at random (MCAR)** when the missingness mechanism is completely independent of the estimate of our parameter(s) of interest.

- **Missing at random (MAR)** when the missingness mechanism is *conditionally* independent of the estimate of our parameter(s) of interest

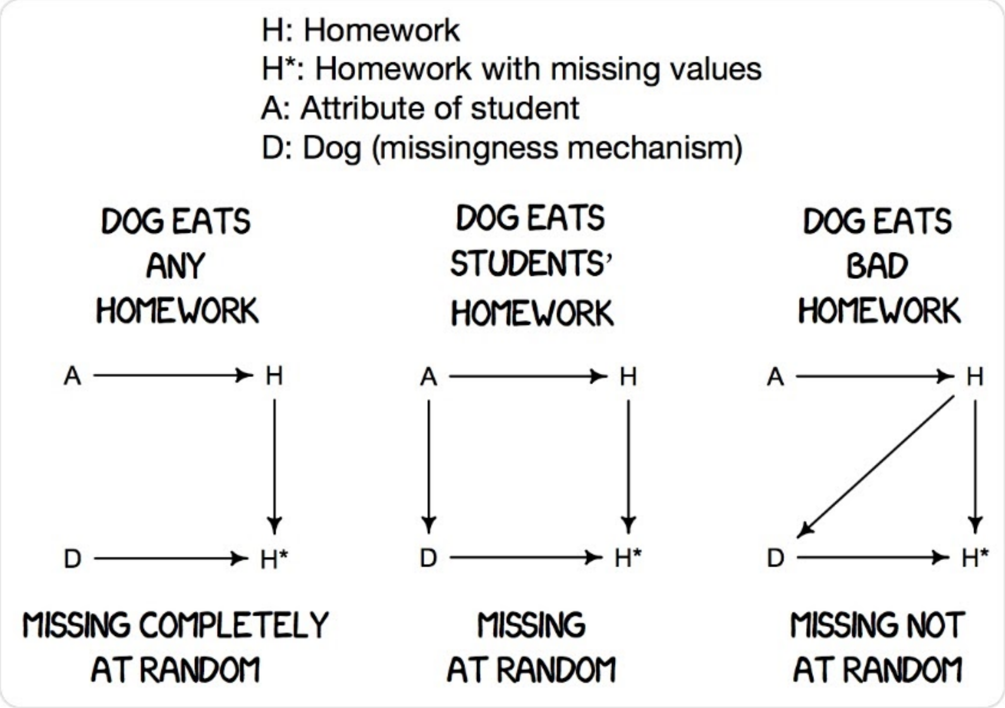
- **Missing not at random (MNAR)** when the missingness mechanism is associated with the estimate of our parameter(s) of interest

Missing Data Mechanisms

Richard McElreath
@rlmcelreath

Follow

In today's lecture, I tried to redefine missing data types (MCAR, MAR, MNAR) as different reasons a dog might eat your homework. This needs more work, but audience seemed to appreciate it.

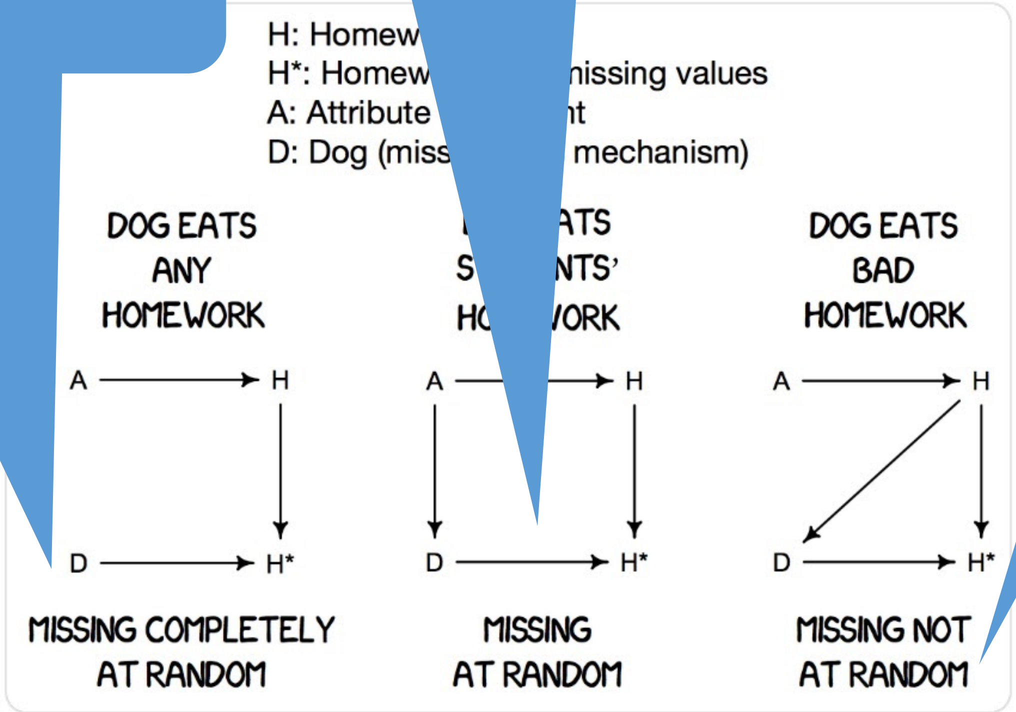


Missing Data Mechanisms

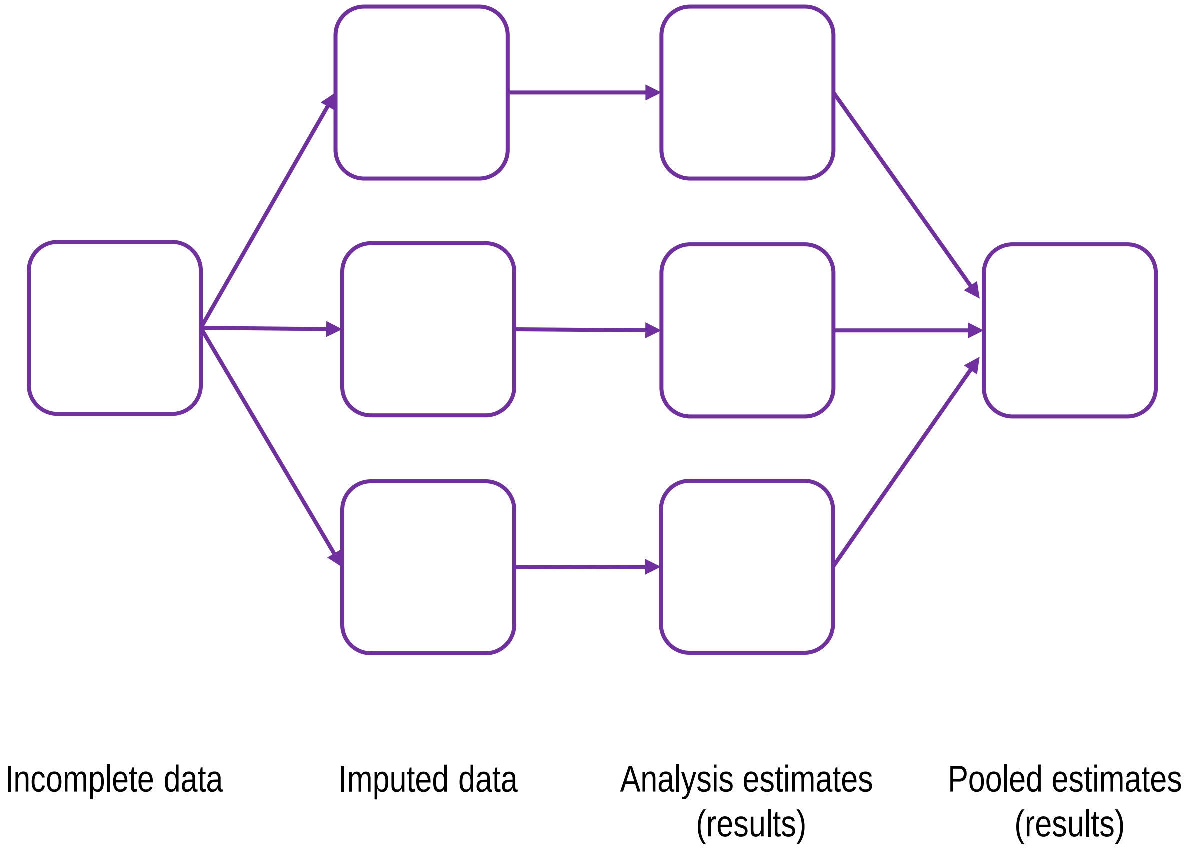
possible to recover unbiased estimates if the right other variables are present.

cannot recover unbiased estimates

listwise deletion will yield unbiased estimates of the true parameter(s) if the data had not been missing



Multiple Imputation



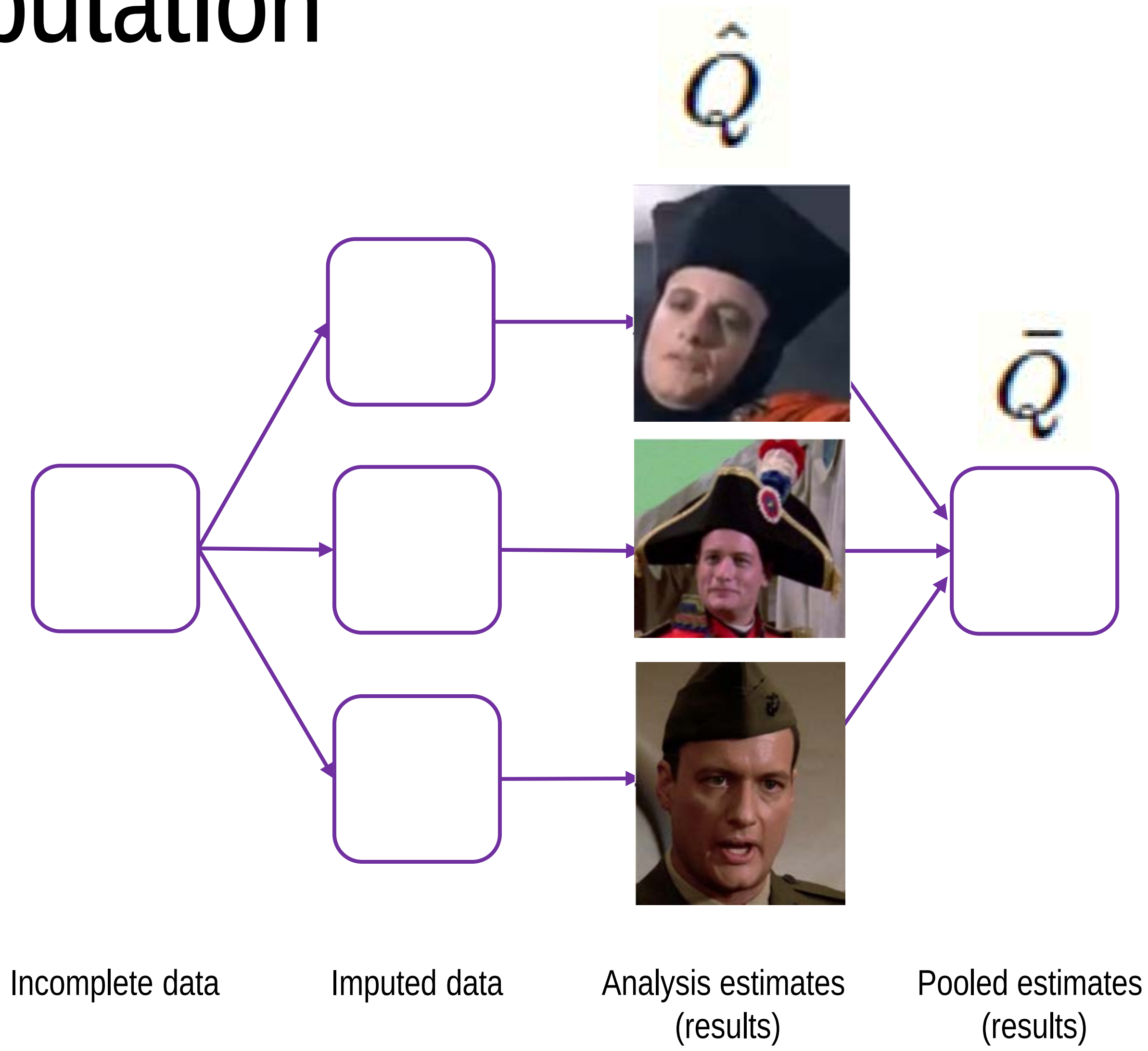
For more info on the MI workflow check out: https://stefvanbuuren.name/fimd/workflow.html

Multiple Imputation

- Let Q be some population value (e.g., a mean, a regression coefficient).
- Let $\hat{Q}$ be an estimate of Q with some estimate of uncertainty due to sampling variation, calculated typically in each imputed dataset.
- Let $\bar{Q}$ be the average of a set of estimates, $\hat{Q}$ across different imputed datasets, with some estimate of uncertainty both due to sampling variation impacting $\hat{Q}$ and missing data uncertainty (causing variation in $\hat{Q}$ from one imputed dataset to the next.

MONASH University

Multiple Imputation



$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i$$

Multiple Imputation

Overall estimate

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$$

Overall variance estimate (uncertainty estimate)

$$\bar{V} = \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i$$

Between-imputed dataset variation:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2$$

Multiple Imputation

Overall variance estimate (uncertainty estimate)

$$\bar{V} = \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i$$

+

Between-imputed dataset variation:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2$$

+

Random sample

$$\frac{B}{m}$$

$$T = \bar{V} + B + \frac{B}{m}$$

Total uncertainty

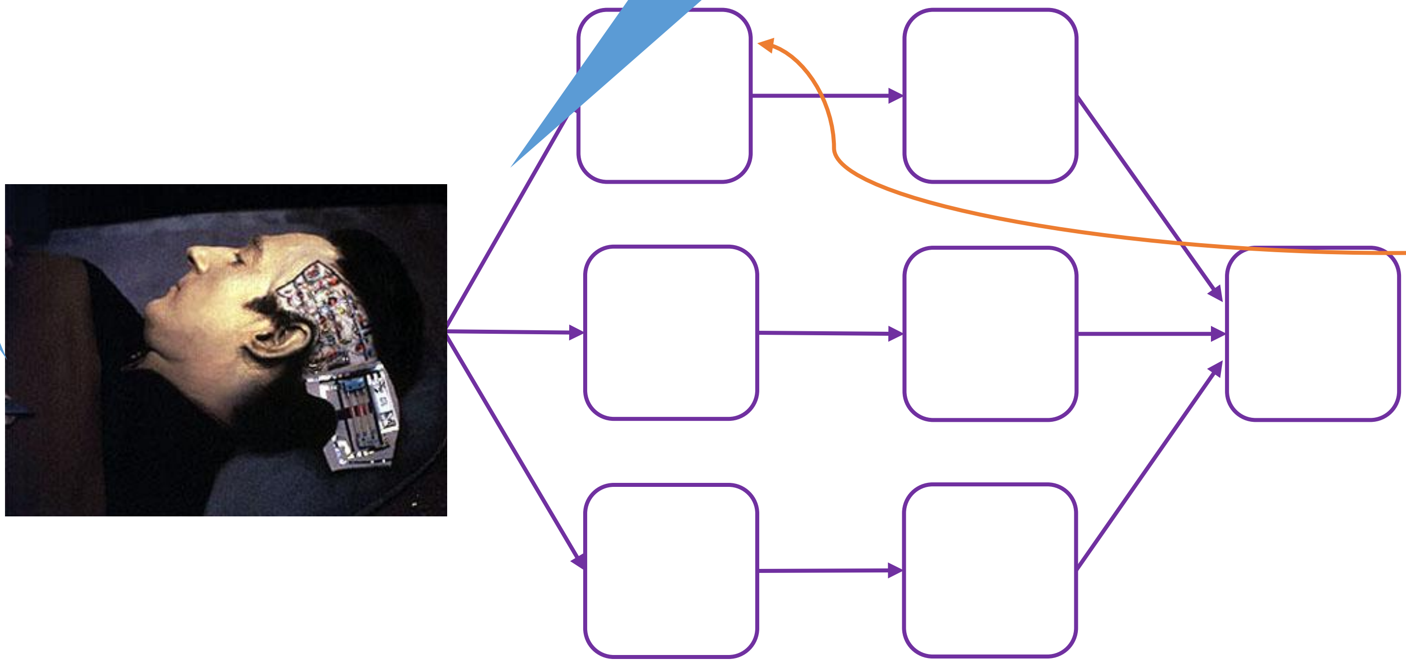
Multiple Imputation

Step 1

Step 2 – very first model uses e.g., means

Step 3 – build prediction model

Step 4 – predict the missing data (if it doesn't vary much, stop repeating 2 & 3, i.e., convergence)



Incomplete data

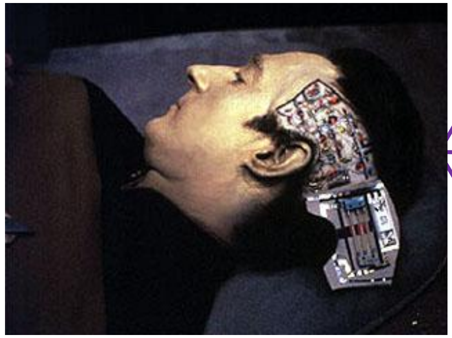
Imputed data

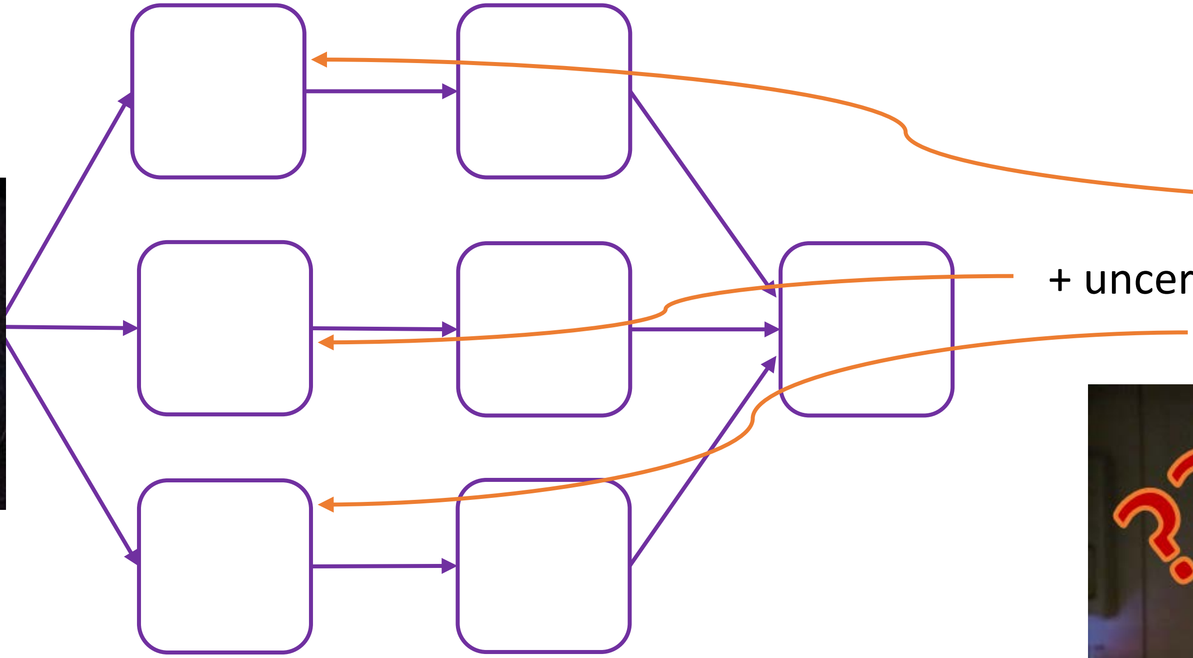
Analysis estimates (results)

Pooled estimates (results)

Multiple Imputation

Step 4 – predict the missing data (if it doesn't vary much, stop repeating 2 & 3, i.e., convergence)





+ uncertainty

Incomplete data

Imputed data

Analysis estimates (results)

Pooled estimates (results)

MONASH University

Prediction models with GLMs

- GLMs assume linear relationships on the scale of the link function.
- GLMs only include interactions between variables when specified by the analyst.
- In small datasets (e.g., 100 people) it is easily possible there may be more variables then people. GLMs require that the sample size be larger then the number of predictors, making them a poor choice for MI in these cases.

smol spocc

MONASH University

Sensitivity Analyses

Developmental Cognitive Neuroscience 33 (2018) 83–98

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

journal homepage: www.elsevier.com/locate/dcn

ELSEVIER

Making an unknown unknown a known unknown: Missing data in longitudinal neuroimaging studies

Tyler H. Matta[a,*], John C. Flournoy[b], Michelle L. Byrne[b]

- Assume MCAR: brain activity during task has no relationship to missing a scan or exclusion due to quality of image (assumption we make when we use listwise deletion)
- MAR: assume a participant's probability to missing a scan is related to brain activity data we DO have
- MNAR: Assume participant's probability of missing a scan related to brain activity data we DIDN'T observe (missing).
- Sensitivity = difference in estimates of the brain activity in complete-case analysis vs. all available data.

Sensitivity Analyses

T.H. Matta, et al. Developmental Cognitive Neuroscience 33 (2018) 83–98

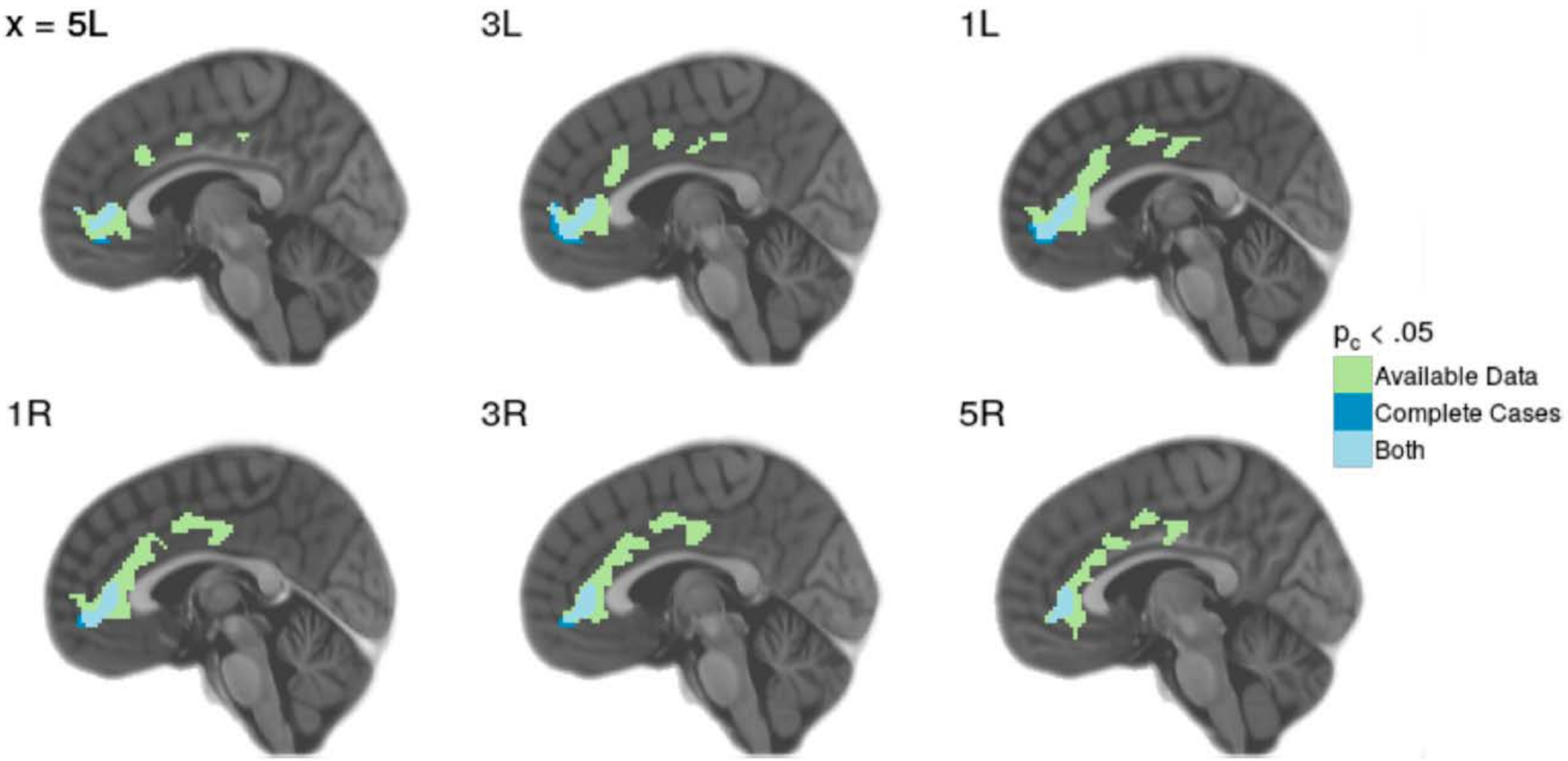


Fig. 3. Significant clusters identified in both available data and complete case analysis is indicated in blue, while significant clusters identified in the available data analysis only are indicated in green. Slice labels indicate the MNI coordinate along which the slice was acquired. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

MONASH University