MICHELLE BYRNE

Senior Lecturer
Monash University

michelle.byrne@monash.edu

COVID Hot Mess & Best Practices for Missing Data in Longitudinal Studies

ABCD Workshop on Brain Development and Mental Health

Sponsored by NIMH

June 22nd - July 22nd, 2021

A lot of this material comes from...

1) Matta, T.H., Flournoy, J.C., Byrne, M.L., 2018. Making an unknown unknown a known unknown: Missing data in longitudinal neuroimaging studies, *DCN*

   https://pubmed.ncbi.nlm.nih.gov/29129673/

   (actually check out our whole special issue: Methodological Challenges in Developmental Neuroimaging: Contemporary Approaches and Solutions https://www.sciencedirect.com/journal/developmental-cognitive-neuroscience/vol/33/suppl/C)

2) Dr. Josh Wiley at Monash University

   Joshuawiley.com

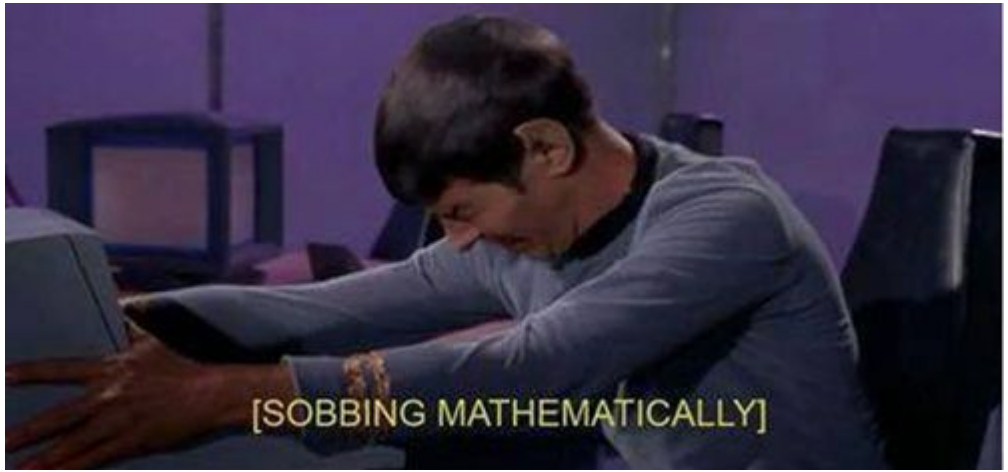
   https://github.com/jwiley (check out his JWileyMisc R package for a lot of useful functions)

3) A perspectives piece yet to be published (stay tuned) on more COVID issues in developmental studies - with Jessica Flannery, Kate McLaughlin, Bonnie Nagel, Jennifer Pfeifer, Eva Telzer, John Flournoy, and Valérie Courchesne

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



Why it matters

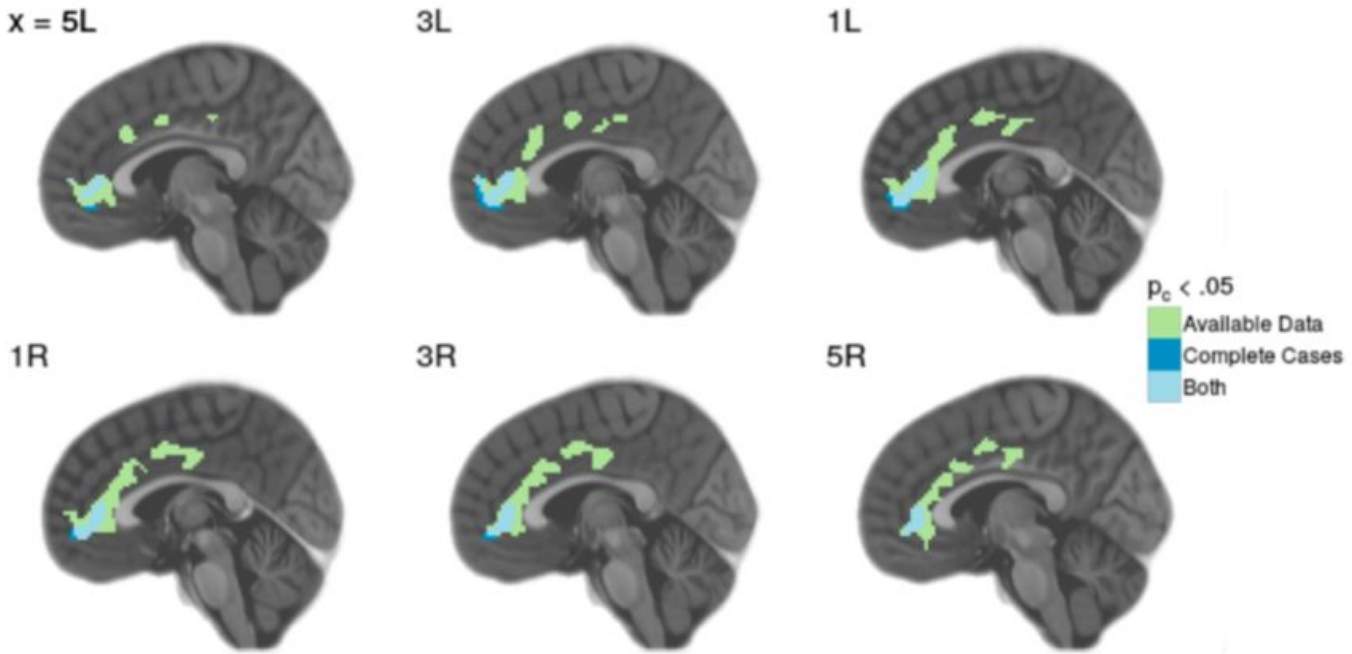


Fig. 3. Significant clusters identified in both available data and complete case analysis is indicated in blue, while significant clusters identified in the available data analysis only are indicated in green. Slice labels indicate the MNI coordinate along which the slice was acquired. (For interpretation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

Some reasons to care about it in ABCD

Baseline brain function in the preadolescents of the ABCD Study

B. Chaarani, S. Hahn, […] the ABCD Consortium

Nature Neuroscience (2021)

https://doi.org/10.1038/s41593-021-00867-9

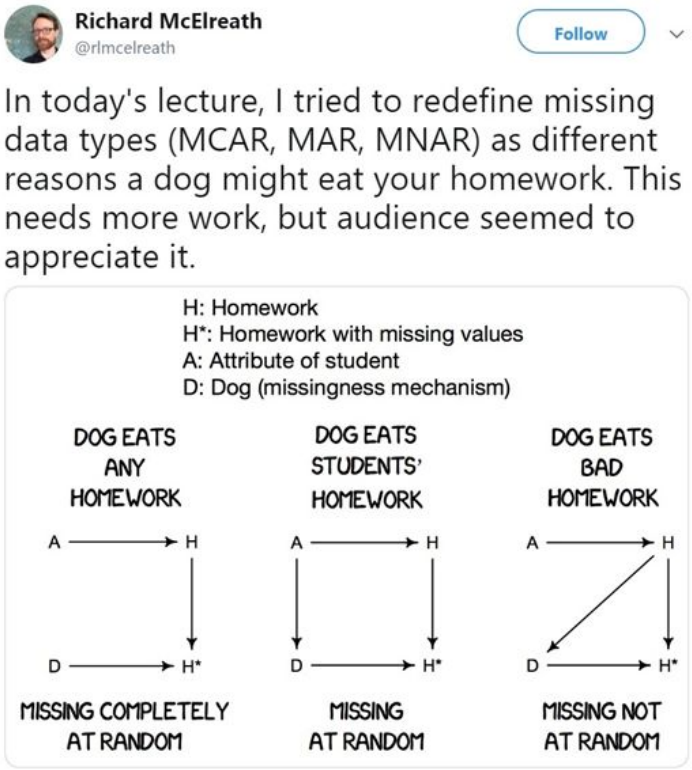
Table 1 | Inclusion criteria in the task fMRI analyses and number of participants remaining after each step of exclusions

Criteria\task	EN-back	SST	MID
Total number scanned participants	10,189	10,294	10,385
Two runs that passed MRI quality control	8,981	9,035	9,140
Data available excluding Philips scans	8,163	8,140	8,201
Mutual vertex and voxel data availability	7,969	7,288	7,427
Motion censoring (mean FD < 0.9 mm)	7,680	7,000	7,239
d.f. across runs >200	7,680	7,000	7,225
Beta weights outlier detection	6,666	6,995	7,214
Passed behavioral performance QC	6,085	5,116	6,753
No missing covariates	6,009	5,547	6,657

Covariates include age, sex, education, puberty, race, family and scanner ID. FD, framewise displacement; QC, quality control.

Missing data type	**Assumptions**	**Conditions for unbiasedness**
Missing completely at random (MCAR): missingness is entirely independent of our outcome of interest (*y*) and any covariates *y* depends on	missingness is (1) not dependent on observations of Y that we don't have (2) not dependent on observations of Y that we do have, and (3) not dependent on covariates X, that Y is dependent on.	Unbiased results using complete case analysis, or all available data in maximum likelihood, multiple imputation.
Covariate-dependent MCAR: missingness is only dependent on the values of variables that affect *y*	(1) and (2); dropping assumption 3.	Unbiased if covariate is included in the models for maximum likelihood analysis or imputation.
Missing at random (MAR): missingness (for a particular participant or unit) is independent of the unobserved values of *y,* i.e., depends only on the values of *y* (for that unit) that we were able to collect	(1); dropping assumptions 2 & 3.	Unbiased estimates only if all available data are used in a maximum likelihood or multiple imputation framework.
Missing not at random (MNAR): missingness is dependent on the unobserved values of *y*, i.e., the values of the missing data depend on the outcome values that we were not able to collect.	No assumptions made.	Biased estimates (sensitivity analyses should be performed).
Take-away: It's best to choose an estimation method that allows you to assume the data are MAR and do sensitivity analyses under the assumption that the data are actually MNAR.		

Missing Data Mechanisms

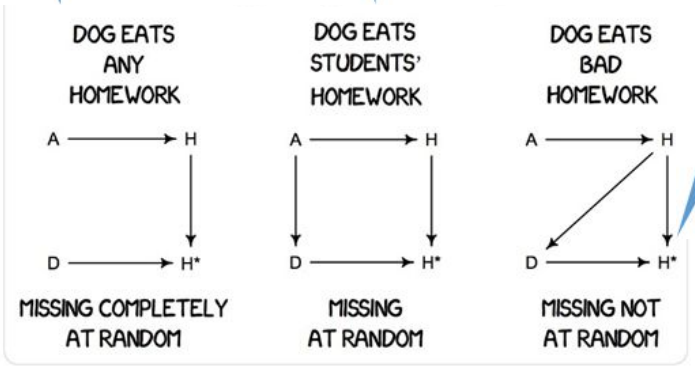


Missing Data Mechanism

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

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

cannot recover unbiased estimates



COVID hot mess

Situation	**Missingness mechanism**	**What do**
Beginning of the pandemic - Research universally shut down - Participants missing an entire wave/MRI	Likely **MCAR** because the temporal order in which participants were assessed should be random with respect to y	missingness depends only on variables that do not affect the outcome; Obtain unbiased estimates using either complete cases or all available data (ML estimation or multiple imputation)
Later in the pandemic - Some regions (or sites!) opened up later - In-person maybe conditional on region	If region was dependent on initial observed anxiety, missingness on an outcome of anxiety would be considered **MAR**	missingness dependent on observed levels of outcome variable; estimate unbiased coefficients using all available data
- The outcome may be influenced by the missingness itself	If missingness is caused by anxiety specific to the missed observation, the data are **MNAR**	the missed observation of anxiety depends on its unobserved level; obtain biased estimates regardless of method - sensitivity analysis
- Tasks shifted online that are harder for younger children (neurocog tasks)	Differences in *age* cause missing an online assessment, and age also causes neurocog performance: **covariate- dependent MCAR**	Obtain unbiased estimates by including age in the model

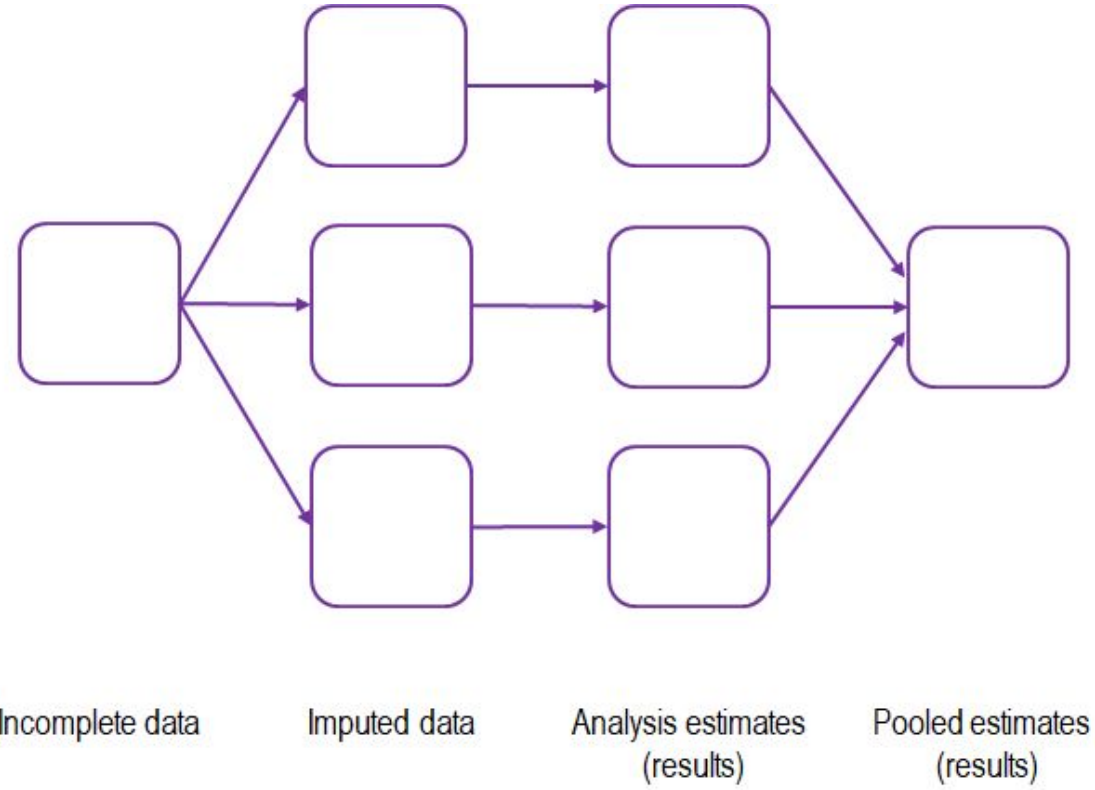
COVID in ABCD

- Offering assessments virtually (over video chat or phone)
- Eliminated some components of the protocol
- Some sites but not others paused in-person visits completely and have only been conducting assessments remotely.
- “Scheduling flexibility” - so some time points might be “off time”.
- Some MRIs will not be completed until the site determines it is safe to reopen
- COVID-19 supplemental data release - more electronic surveys specific to coping during the pandemic and other situational variables (May, June, Aug, Dec 2020).

Recommendations:

1. Always include any variables upon which *y* theoretically (not empirically) depends.
2. Carefully consider the COVID-related missing data mechanism. Is the missing outcome variable conditionally dependent on some other variable that *did* get collected (MAR or covariate-dependent MCAR), or is it assumed that the value of the outcome variable (if we did know it) is the reason for the missingness (MNAR)?
3. It is impossible to empirically rule out MNAR. Sensitivity analyses can be performed to assess the effect of missing observations had they been collected. Tutorials: Coertjens et al., 2017; Leurent et al., 2018; Resseguier et al., 2011.
4. Always conduct analyses with all available data using maximum likelihood or multiple imputation methods. In rare cases, ML methods may not be available in which case multiple-imputation may be helpful.

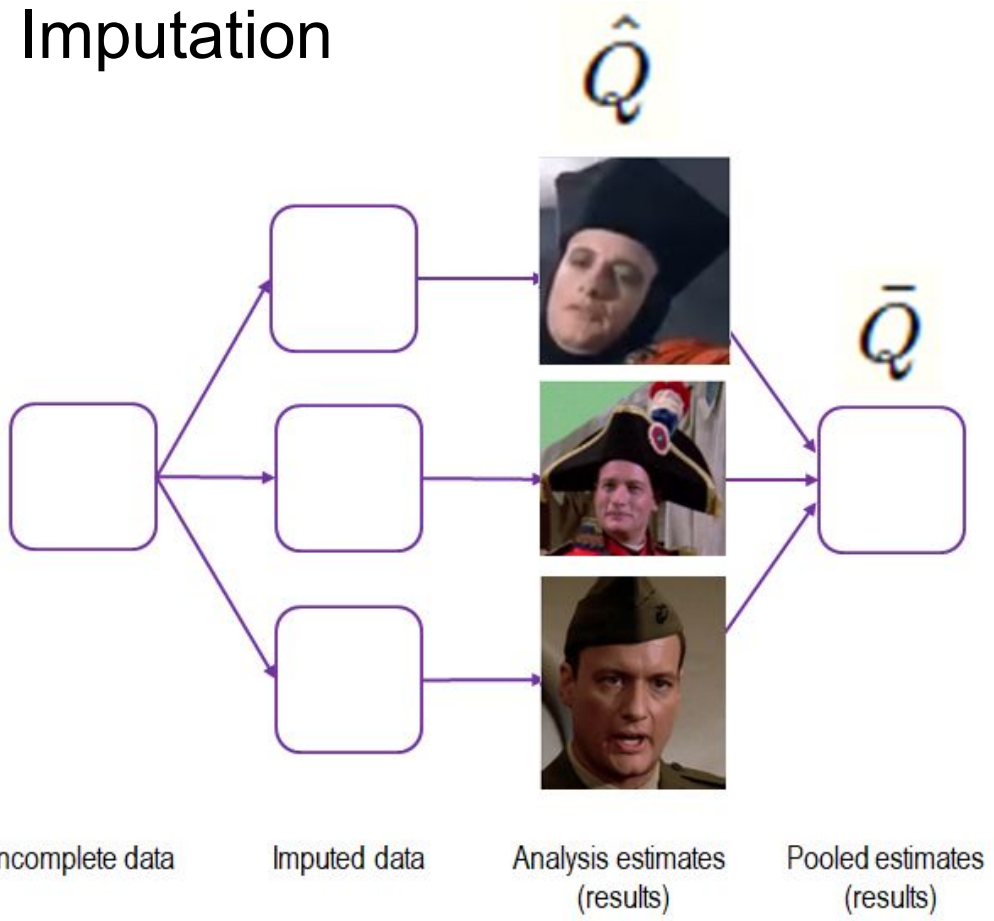
A very crash course in Multiple Imputation (MI)



Multiple Imputation

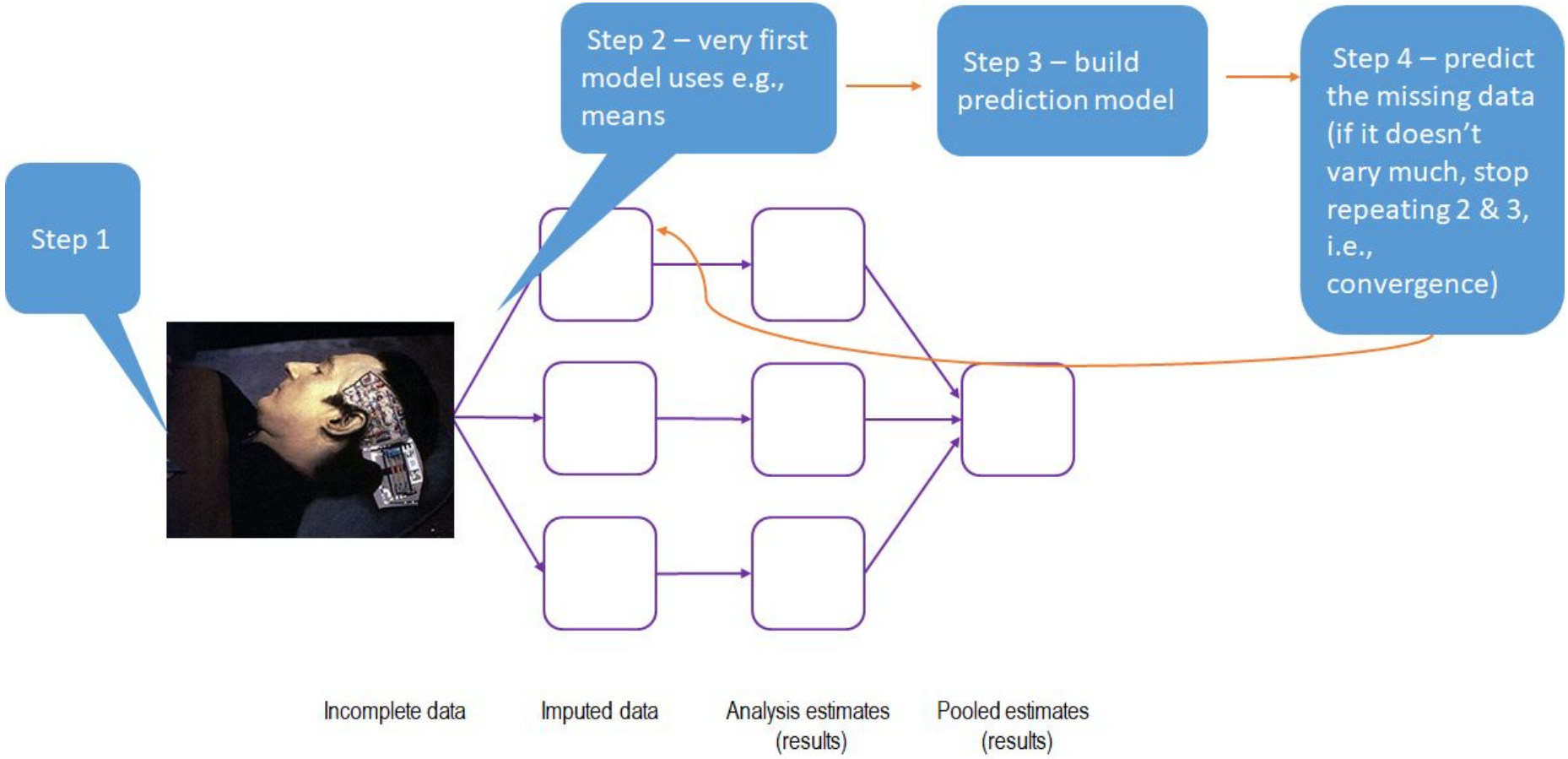
- Let Q be some population value (e.g., a mean, a regression coefficient).
- Let $\hat{Q}$ be an estimate of Q with some estimate of uncertainty due to sampling variation, calculated typically in each imputed dataset.
- Let $\bar{Q}$ be the average of a set of estimates, $\hat{Q}$ across different imputed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

Multiple Imputation

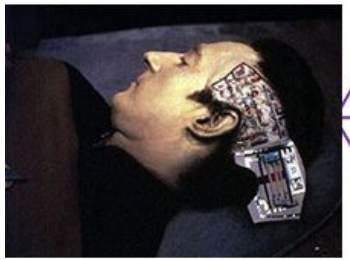


$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$$

Multiple imputation STEPS



Multiple imputation STEPS



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

+ uncertainty

Neat code from Josh Wiley to show how imputation works

https://github.com/michellebyrne1/MonashHonoursStatistics/blob/8e16b6f141340869bb99296aa25940dafe6cdd7a/MissingData_Michelle.Rmd#L233