Evaluation of approaches for multiple imputation of three-level data

Rushani Wijesuriya
Dr. Margarita Moreno-Betancur, Prof. John Carlin, Dr. Anurika De Silva and Prof. Katherine Lee,

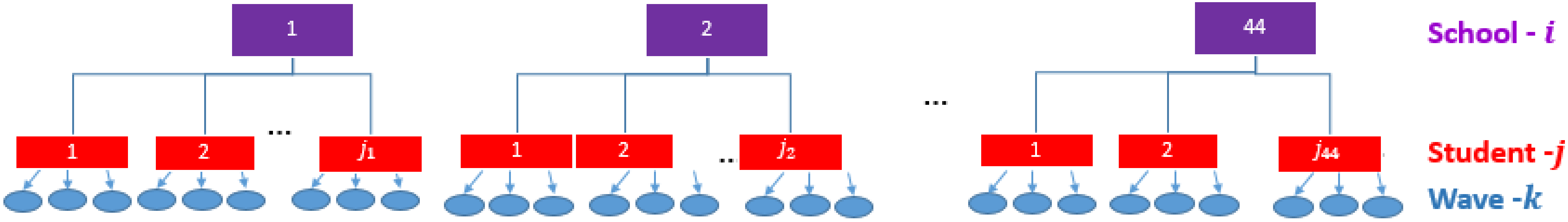
The Clinical Epidemiology and Biostatistics Unit(CEBU)
University of Melbourne

Australian and New Zealand Statistical Conference 2021

Overview

- The motivating case study
- Multiple imputation
- Research aim
- Simulation study
- Conclusions

The motivating case study: Childhood to Adolescence Transition Study (CATS)



Three levels of hierarchy

- Repeated measures within an individual and also clustering by school

Mundy, L. K., et al. (2013). "Study protocol: the childhood to adolescence transition study (CATS)." BMC pediatrics **13**(1): 160.

The motivating case study : Target analysis and missing data

- Substantive research question :

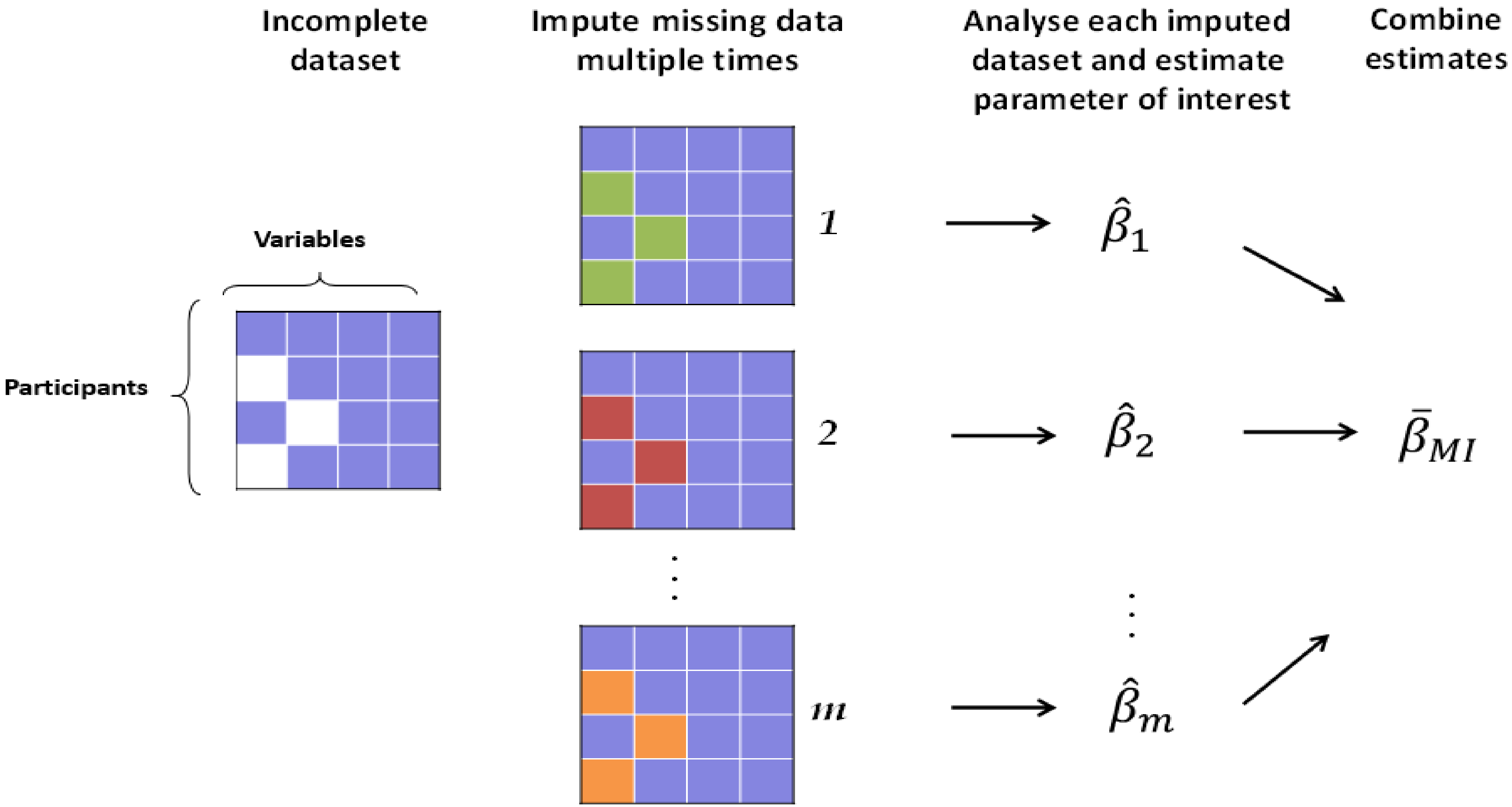
The effect of early depressive symptoms on the academic performance of the students

$$\boxed{NAPLAN_{ijk}} = \beta_0 + \beta_1 \times \boxed{depression_{ij(k-1)}} + \beta_2 \times wave$$
$$+ \beta_3 \times \boxed{NAPLAN_{ij1}} + \beta_4 \times sex_{ij} + \beta_5 \times SES_{ij1} + \beta_6 \times age_{ij1}$$
$$+ b_{0i} + b_{0ij} + \varepsilon_{ijk}$$

Where i denotes the i^{th}school, j denotes the j^{th} individual and k denotes the k^{th} wave

$i = 1, \dots, 44$
$8 \leq j \leq 66$
$k = 3,5,7$

Multiple Imputation



Common implementation frameworks :

- Joint modelling (JM)
- Fully conditional specification (FCS)

Figure designed by Nguyen C: Diagnostic methods for checking multiple imputation models. Parkville: The University of Melbourne; 2014

Multiple Imputation

- Imputation model specification in MI:

  Congeniality between the imputation and analysis model(1)

Need to incorporate important features of the analysis (such as clustered structures, interactions, non-linearities etc) in the imputation model

(1)Meng, X.-L. (1994). "Multiple-imputation inferences with uncongenial sources of input." Statistical Science: 538-558.

Multiple imputation of two-level data

Adaptations of standard (single-level) MI approaches

- For cluster groups (such as schools)- Dummy Indicator (DI) approach
- For repeated measures (at fixed intervals)- Impute in wide format

Wide format
one row per individual

ID	Age	Sex	Dep_1	Dep_2	Dep_3
1	8	Male	0.4	1.9	0.2
2	7	Female	1.9	-	2.9
3	9	Male	1.0	3.1	-
4	8	Male	-	2.6	-
5	10	Female	1.5	0.5	1.5

MI approaches based on (two-level) mixed effects models(1)

- Both JM and FCS implementations are available
- More recent than single-level implementations

Long format
One row per wave per individual

ID	Age	Sex	Wave	Dep
1	8	Male	1	0.4
1	8	Male	2	1.9
1	8	Male	3	0.2
2	7	Female	1	1.9
2	7	Female	2	-
2	7	Female	3	2.9

(1)Schafer, J. L. and R. M. Yucel (2002). "Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values." Journal of Computational and Graphical Statistics **11**(2): 437-457.

Multiple Imputation of three-level data

- Adaptations of the single-level MI methods
  - For cluster groups : Dummy indicator (DI) approach
  - For repeated measures (at fixed intervals): Impute in wide format
  - JM-1L-DI-wide
  - FCS-1L-DI-wide

- Adaptations of MI approaches based on two-level (RE) models
  - For cluster groups: DI approach
  - For repeated measures: Two-level MI approach (RE)
  - JM-2L-DI
  - FCS-2L-DI

  - For cluster groups: Two-level MI approach (RE)
  - For repeated measures: Impute in wide format
  - JM-2L-wide
  - FCS-2L-wide

- MI approaches based on three-level (RE) models
  - For cluster groups: RE
  - For repeated measures: RE
  - FCS-3L

Aim

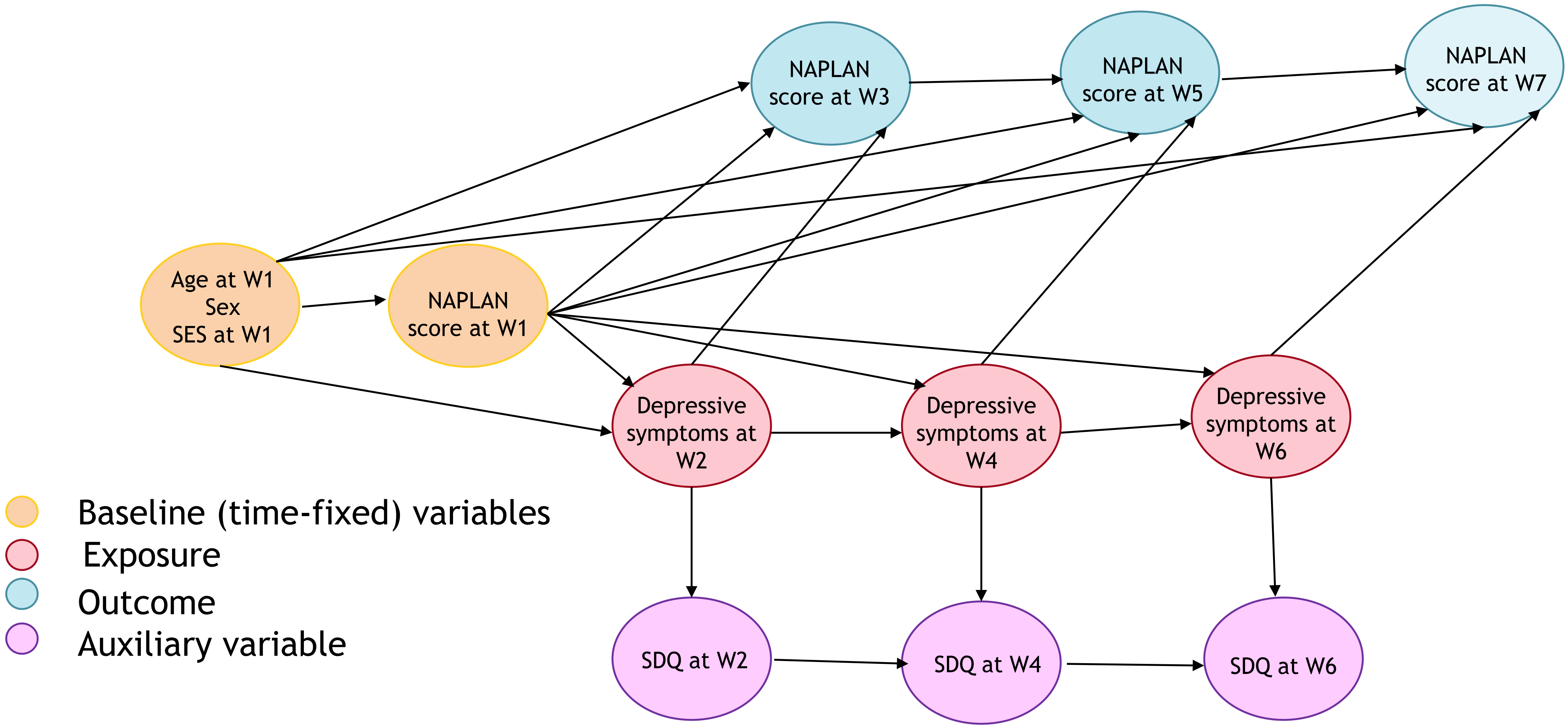
- Evaluate MI approaches for handling incomplete three-level data using a simulation study with the aim of providing guidance on the use of these approaches

Simulation of Complete Data

- Mimicking the real CATS data
- 40 school clusters $(i = 1, \ldots, 40)$ were generated
- Each school cluster was populated
- Four different strengths of level-2 and level-3 intra-cluster correlations:

	ICC	
	level 3 (within school)	level 2 (within individual)
High-high	0.15	0.5
High-low	0.15	0.2
Low-high	0.05	0.5
Low-low	0.05	0.2

Simulation of complete data



Generation of missing data

MCAR

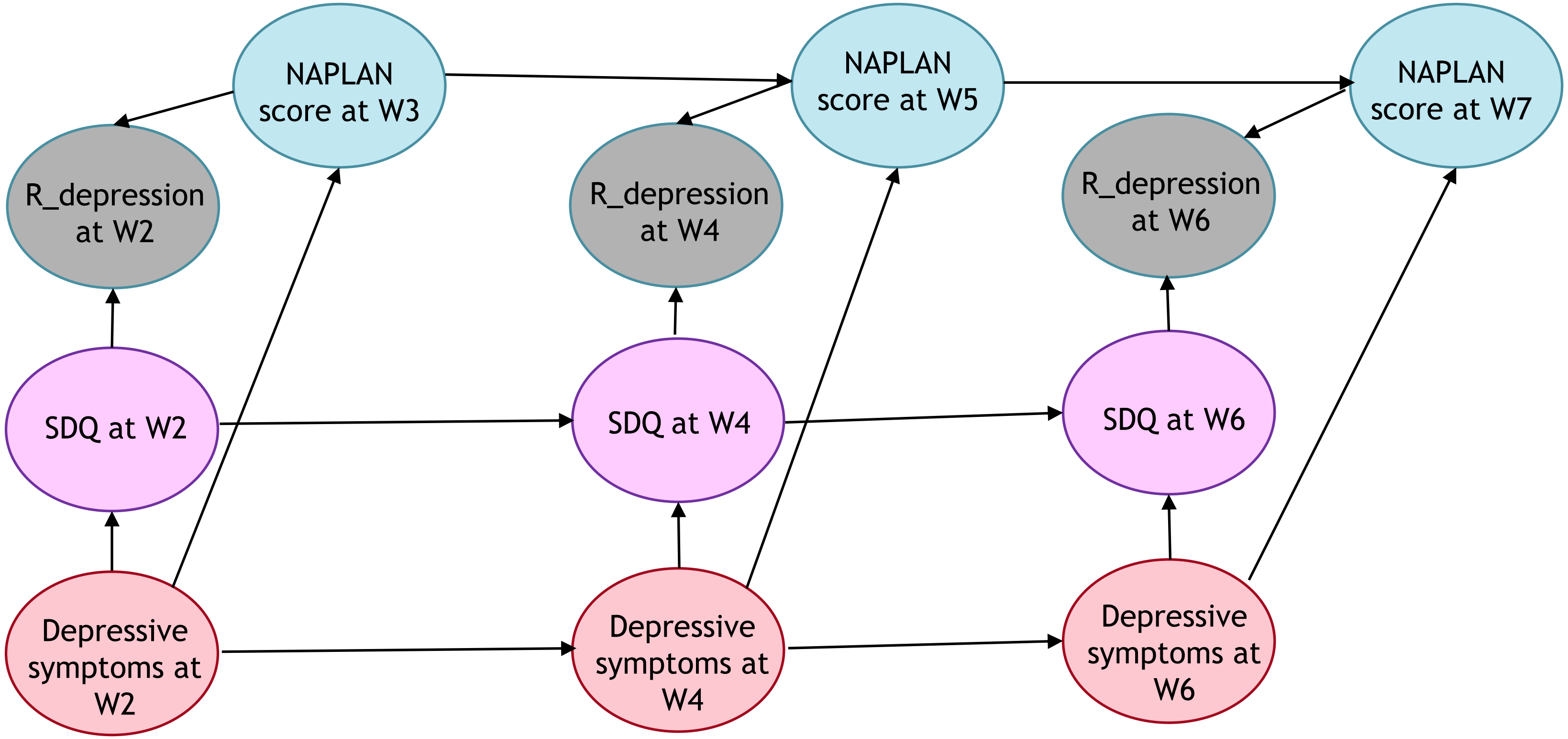
Missing values assigned completely at random

10%	Depressive symptoms at W2	20%
15%	Depressive symptoms at W4	30%
20%	Depressive symptoms at W6	40%

MAR

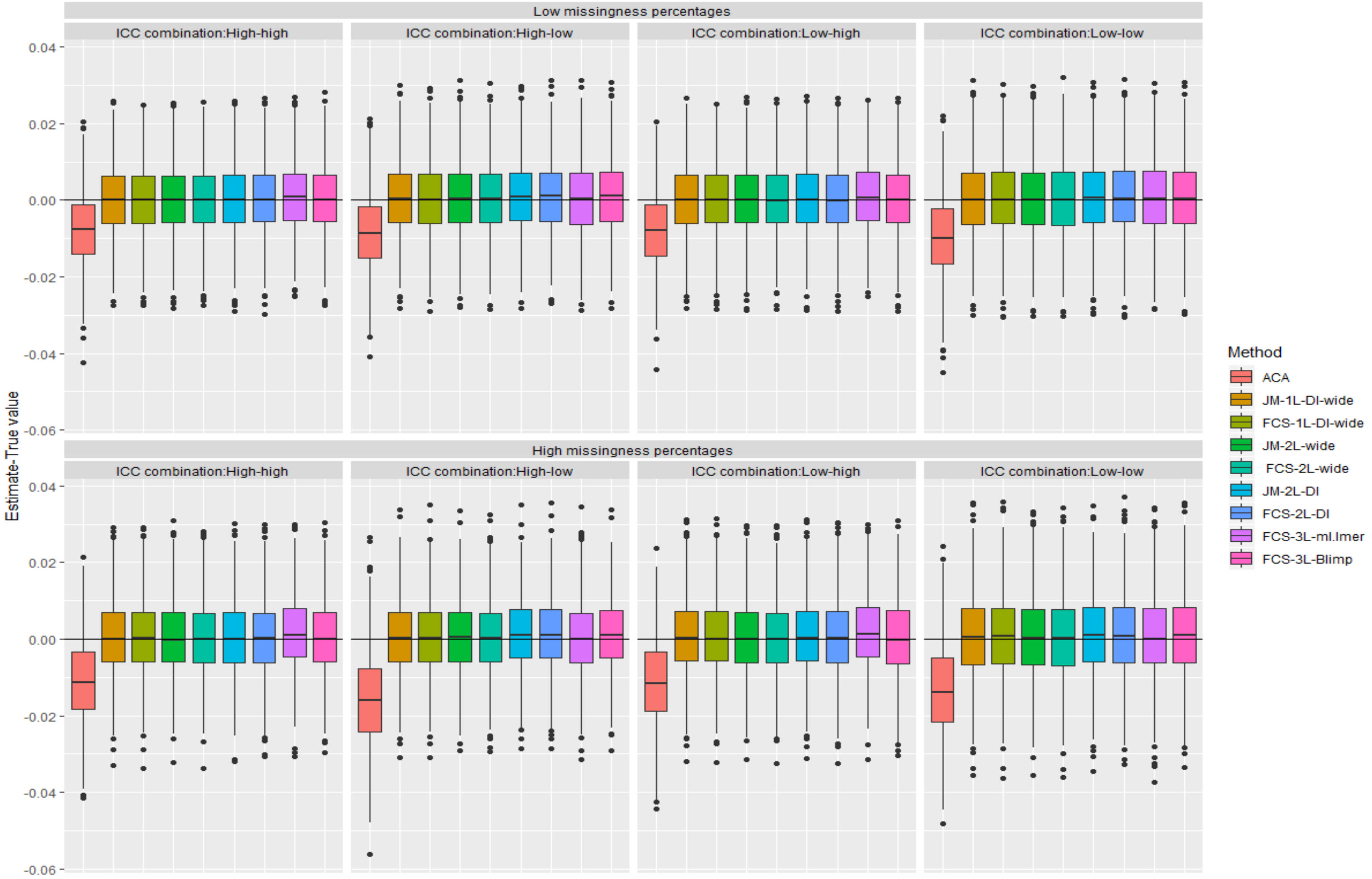
Similar to CATS (MAR-CATS)

Strong (MAR-inflated)



Simulation Study-Results

$\beta_1 = (-0.025)$



Results shown only for MAR-CATS scenario as the comparative performance of approaches was similar under MAR-inflated scenario

Conclusions

- All approaches can be used to handle incomplete three-level data in the context of a random intercept substantive analysis model
- Approaches which use the DI extension should be used with caution
- In the presence of longitudinal data measured at irregular time intervals, three-level imputation approaches will be required

Thank you

Find me at:

https://www.rwijesuriya.com/

@rush_099

@rushwije

Scan to access published article of this work

murdoch children's research institute

THE UNIVERSITY OF MELBOURNE

Statistical Society of Australia

ViCBiostat