Multiple imputation of missing values for randomized controlled trials: A step-by-step tutorial using *mice*

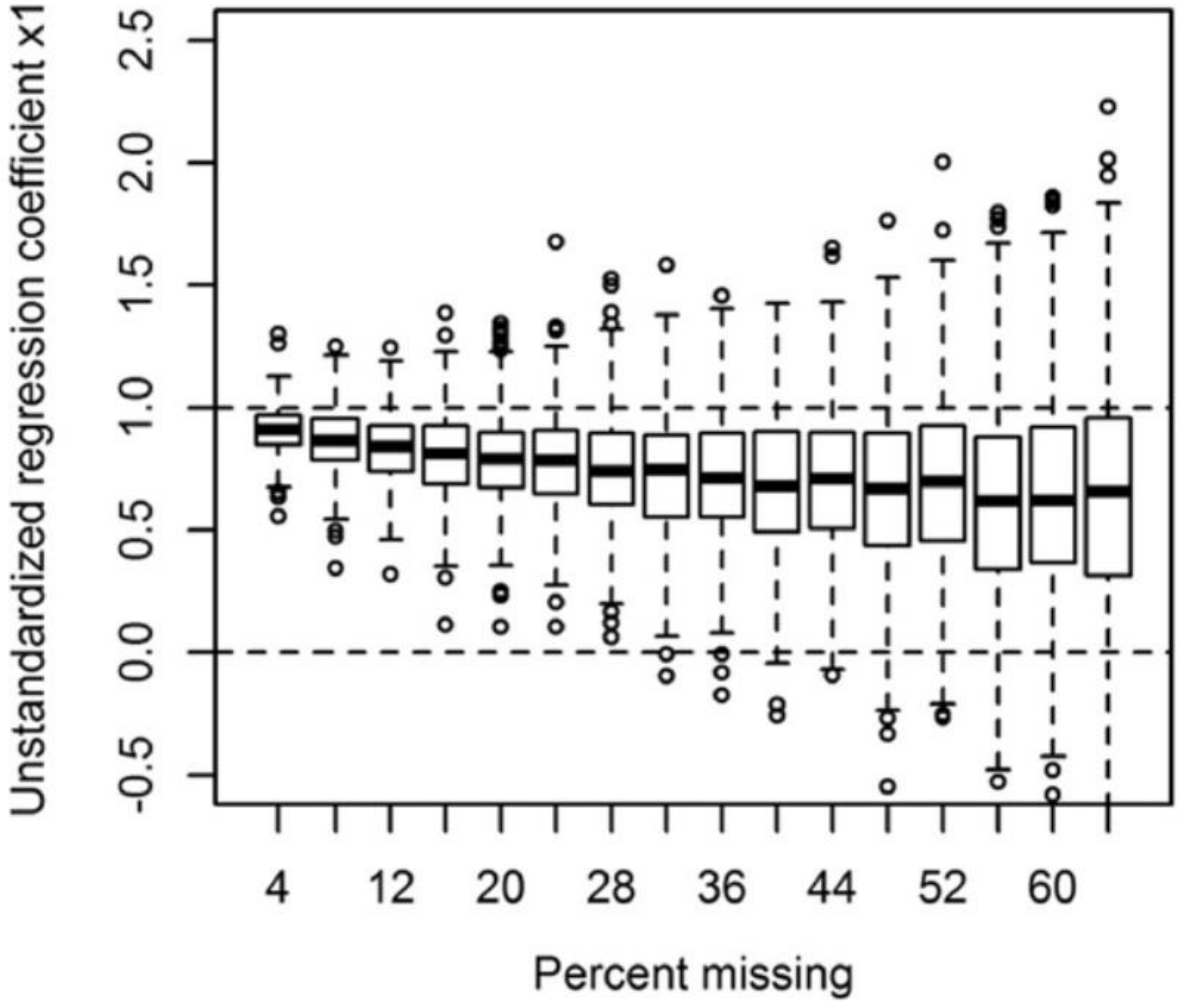
Oscar Lecuona, Ariadna Angulo-Brunet, Víctor Ciudad, Ricardo Olmos

olecuona@ucm.es

Introduction

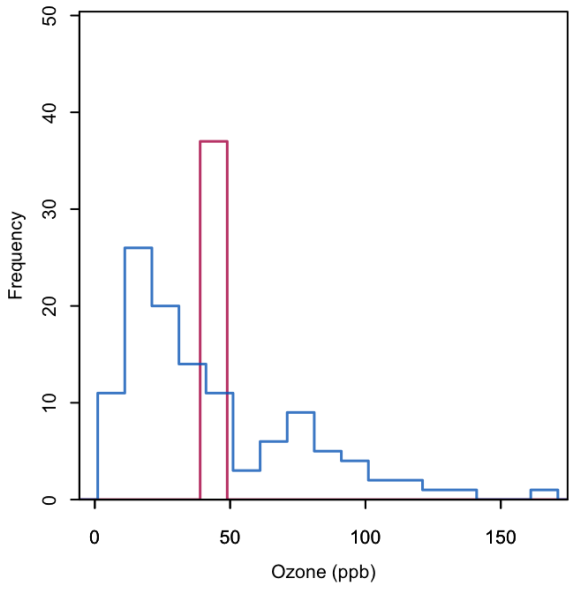
- Randomized Controlled Trials are a gold standard to test effectiveness of interventions
  - Primary source for meta-analysis
- Missing data is common in RCTs
  - Patient dropout (MAR or MNAR)
  - Random missingness (e.g., accidentally not responding, MCAR)
- Missingness is a relevant issue:
  - Power decreases
  - Bias increases
  - More likely to get technical issues

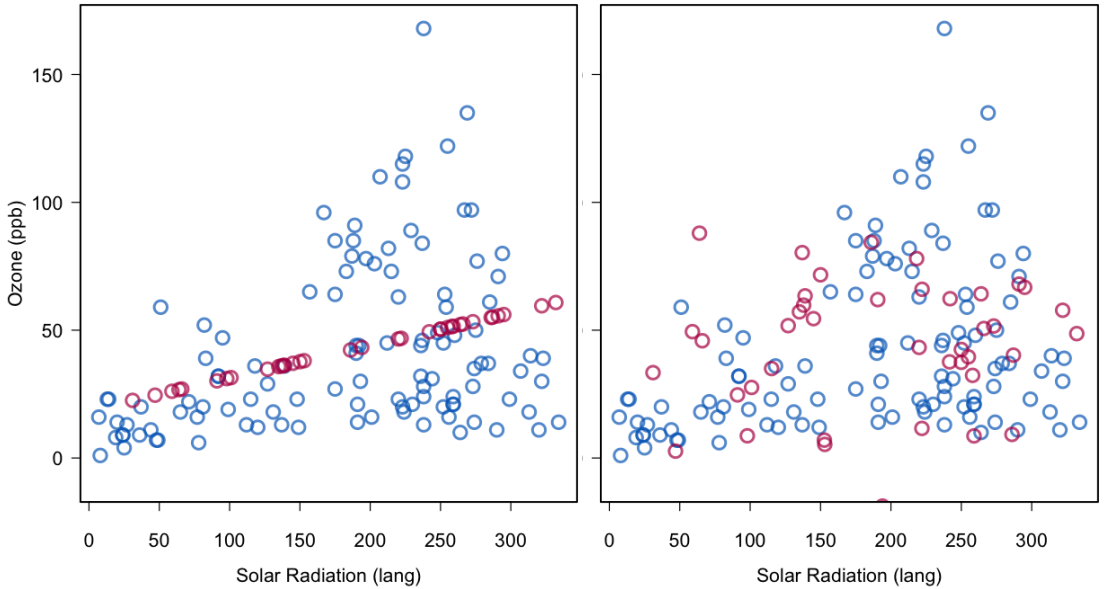
http://www.biomedcentral.com/1471-2288/12/184

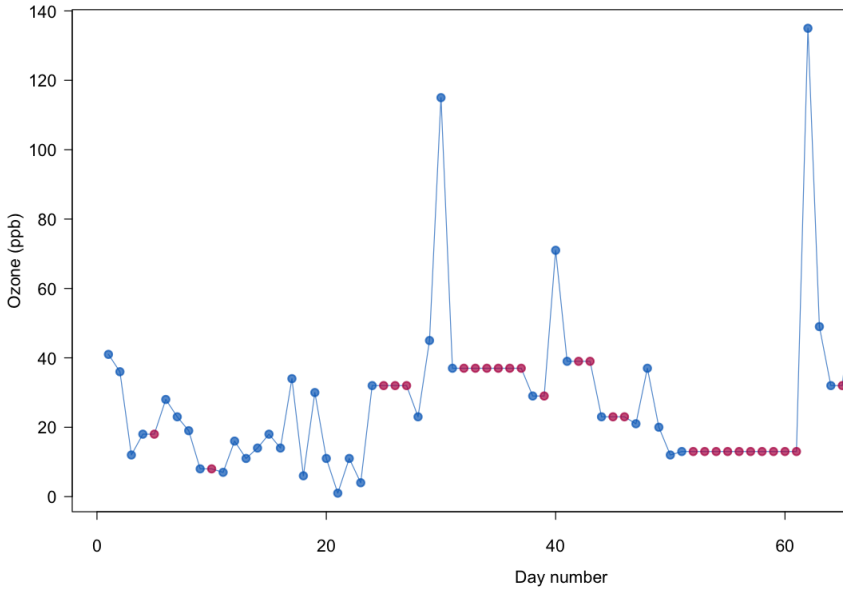


What we can do?

- All analysis techniques require one
- Common techniques:
  - Complete Case Analysis (listwise deletion, pairwise deletion)
  - Simple imputation
    - Mean, median, regression, stochastic methods
    - Last-Observation-Carried-Forward (LOCF), Baseline-Observation-Carried-Forward (BOCF)







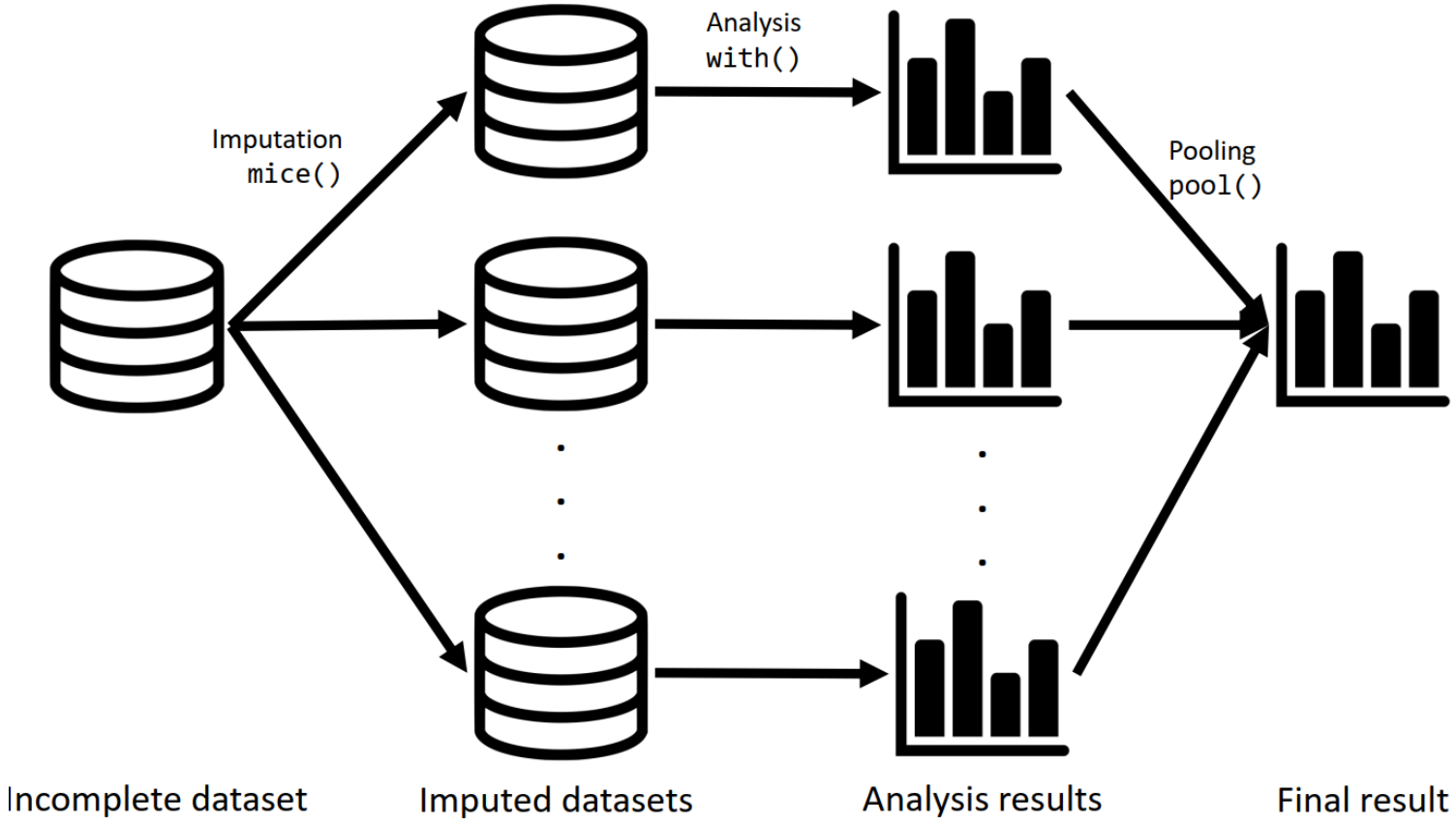
What we can do?

- Full-Information Maximum Likelihood (FIML)
  - Requires specifying a model from theory, multivariate normality, and linearity
    - *Multivariate Normal Imputation (MVNI)*
- Bayesian Imputation (BI)
  - Requires specifying a chosen model from theory
  - Treat missing values as latent variables
- Multiple imputation (MI; e.g., *mice*)
  - Model is specified with present data; open to all types of data ("agnostic")
    - *Joint modelling (JM), fully conditional specification (FCS)*

	FIML	BI	MI
Provides unbiased estimates	Yes	Yes (if prior is appropriate)	Not always (data-dependent)
Requires theoretical model	Yes (continuous, normal, linear)	Yes (specified by researchers)	No (taken from data)
Works with non-linear, non-normal, mixtures	"No" (in principle)	Yes (if specified)	Yes (all kinds of data)
Multilevel and interactions	"No" (in principle)	Yes (specified)	Yes (specified)
Available for most techniques	No (mainly SEM)	No	Yes

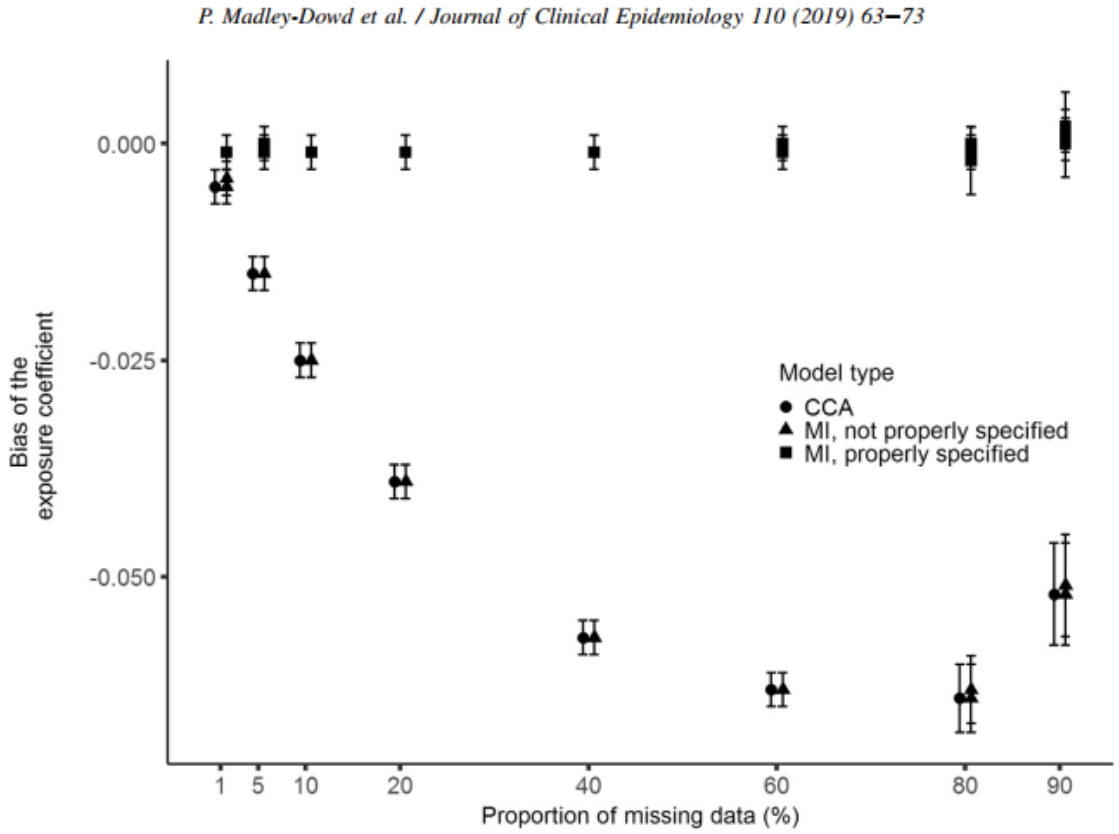
Multiple imputation with *mice*

- Incomplete data is screened (MCAR, MAR, MNAR) to find patterns and select:
  - Which variables will predict each variable with missing values
  - Estimation method (parametric, non-parametric, etc.)
- Markov Chain Monte Carlo simulations create n imputed datasets
  - Each one produces one result (e.g., ANOVA)
- Results are combined taking into account variance between datasets



Multiple imputation with *mice*

- If properly specified:
  - Bias reduction + power increase
  - If not, bias can be the same than CCA
- What to specify?
  1. Data structure (1 level, 2-level, 3-level, etc.)
  2. Predictors:
     - Theoretical criteria: In RCTs, group and time
     - Empirical criteria: variables correlated >.3 to variables with missingness
     - Care for too many (can bias results)
  3. Estimation method:
     - Generalist: *pmm*, *sample, pan, jomo,…*
     - Specialized: *norm, logistic, polyreg,…*
  4. Samples/Chains (how many)
     - 3 to 5, mean% of missing data, or “fraction of missing information” (FMI)
  5. Iterations (how many)
     - Min. 50, higher = better, inspect afterwards



https://doi.org/10.1186/1742-7622-9-3

Why a tutorial on mice for RCTs?

- MI is a complex but demanded technique
  - Main need: Dropout in RCTs
- Current tutorial literature on MI seems improvable in offering concrete and step-by-step solutions
- This tutorial is meant for advenced users
  - Advice for applied users to work with an expert or self-train

Structure of the tutorial

- Data screening
- Missingness assessment
- Building the imputation model
- Estimating the model
- Results assessments
- Sensitivity Analysis

Script and data here:

Example database

- Randomized Controlled Trial (n = 82)
- Two active programs: a well-established mindfulness intervention (MBCT) and a new program (WBT)
- Four time points: Pre, post, follow-ups (6 months, 1 year)
- Demographics: middle-aged (40), mostly female (90%), single/married (80%), BAs/MDs (80%), no prior experience with meditation (60%)
- Dependent variables: well-being (WS, 1 factor), psychopathology (DASS-21, 3 factors), mindfulness (FFMQ-SF, 5 factors)
- No differences in demographics or pre scores

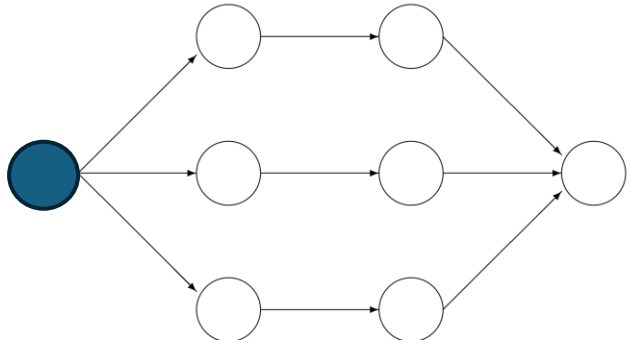
Characteristic	**Overall**, N = 82[1]	**1**, N = 47[1]	**2**, N = 35[1]	**p-value**[2]
W	26.27 (3.64)	26.15 (3.92)	26.43 (3.28)	0.8
DASS_D	8.4 (6.3)	8.1 (5.9)	8.8 (6.8)	0.7
DASS_A	4.2 (4.8)	4.3 (4.1)	4.2 (5.9)	0.6
DASS_E	12 (7)	13 (6)	11 (9)	0.3
FFMQ_OB	9.49 (2.66)	9.72 (2.79)	9.17 (2.48)	0.5
FFMQ_DS	10.77 (2.18)	10.47 (2.20)	11.17 (2.12)	0.083
FFMQ_AC	8.77 (2.14)	8.60 (2.03)	9.00 (2.30)	0.7
FFMQ_NJ	10.79 (2.43)	10.62 (2.55)	11.03 (2.28)	0.4
FFMQ_NR	9.05 (1.85)	9.11 (1.72)	8.97 (2.02)	>0.9

[1] Mean (SD)

[2] Wilcoxon rank sum test

Example database

- Load packages, data, examine
- Turn to long format (time as a variable)



id <chr>	group <chr>	sex <chr>	age <chr>
1	1	1	47
2	1	1	57
3	1	1	48
4	2	1	47
...	...	...	...
79	1	2	49
80	2	1	47
81	2	1	58
82	2	1	47

9 rows | 1-10 of 43 columns

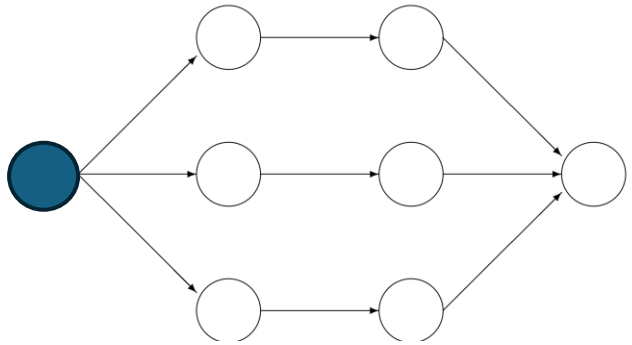
...

T1_W <chr>	T2_W <chr>	T3_W <chr>
26	28	29
27	31	30
30	34	31
28	28	*NA*
...	...	...
24	22	*NA*
25	22	22
24	27	25
23	26	23

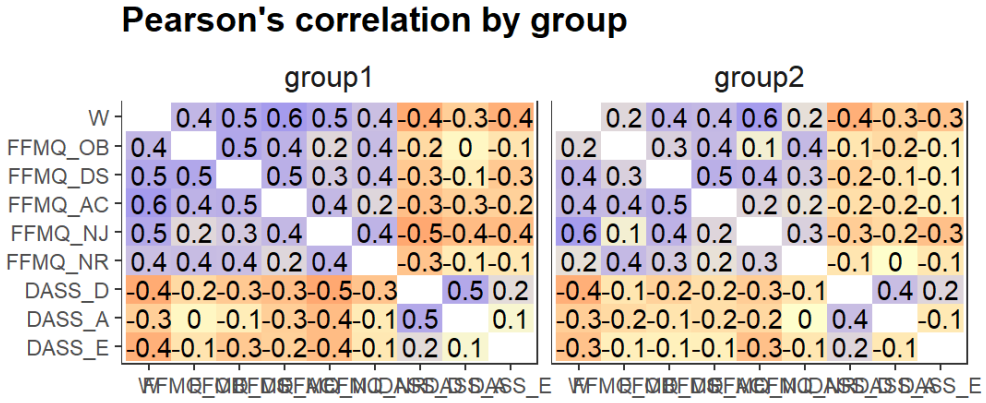
time <chr>	W <dbl>	FFMQ_OB <dbl>	FFMQ_DS <dbl>	FFMQ_AC <dbl>
T1	26	8	8	9
T2	28	10	9	10
T3	29	11	10	11
T4	28	10	9	11
T1	27	13	12	11
T2	31	13	12	11
T3	30	14	13	11
T4	30	12	12	9
T1	30	8	12	12
T2	34	13	12	14

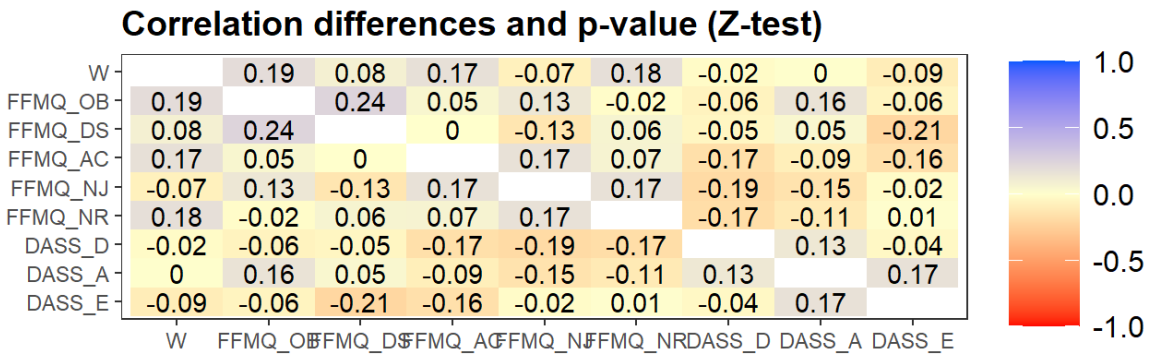
Previous 1 2 3 4 5 6 ... 33 Next

Data Screening

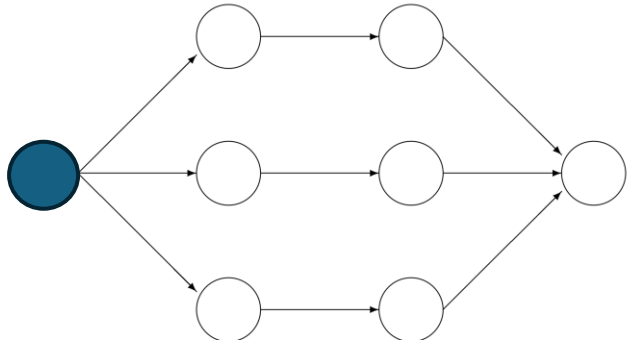


- Data structure: 2-level (time and groups nested in participants/"Id")
- Group: Split the data
  - Allows interactions to take place (3-level not needed)
- Check if correlations between variables differ between groups

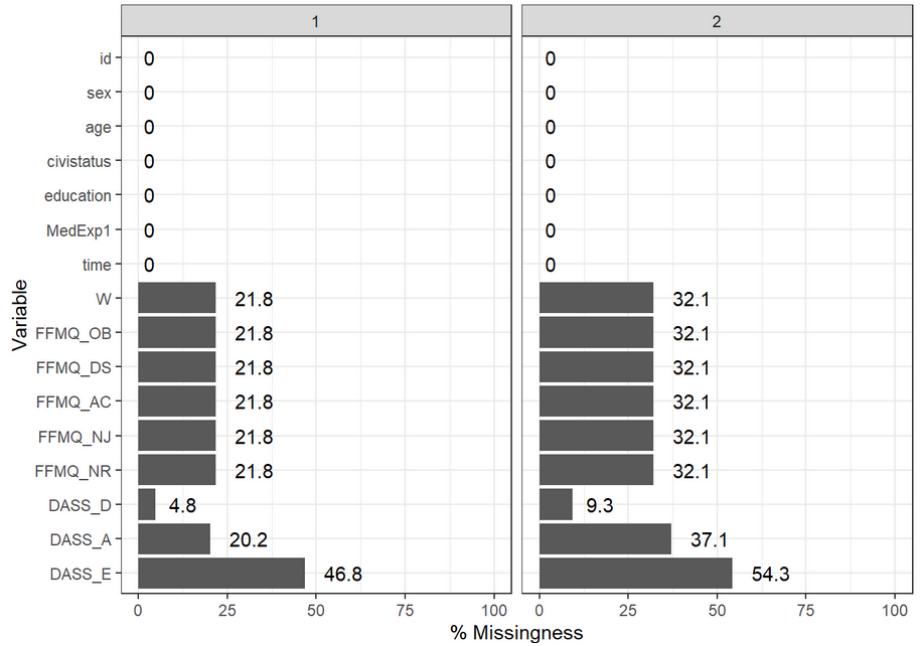


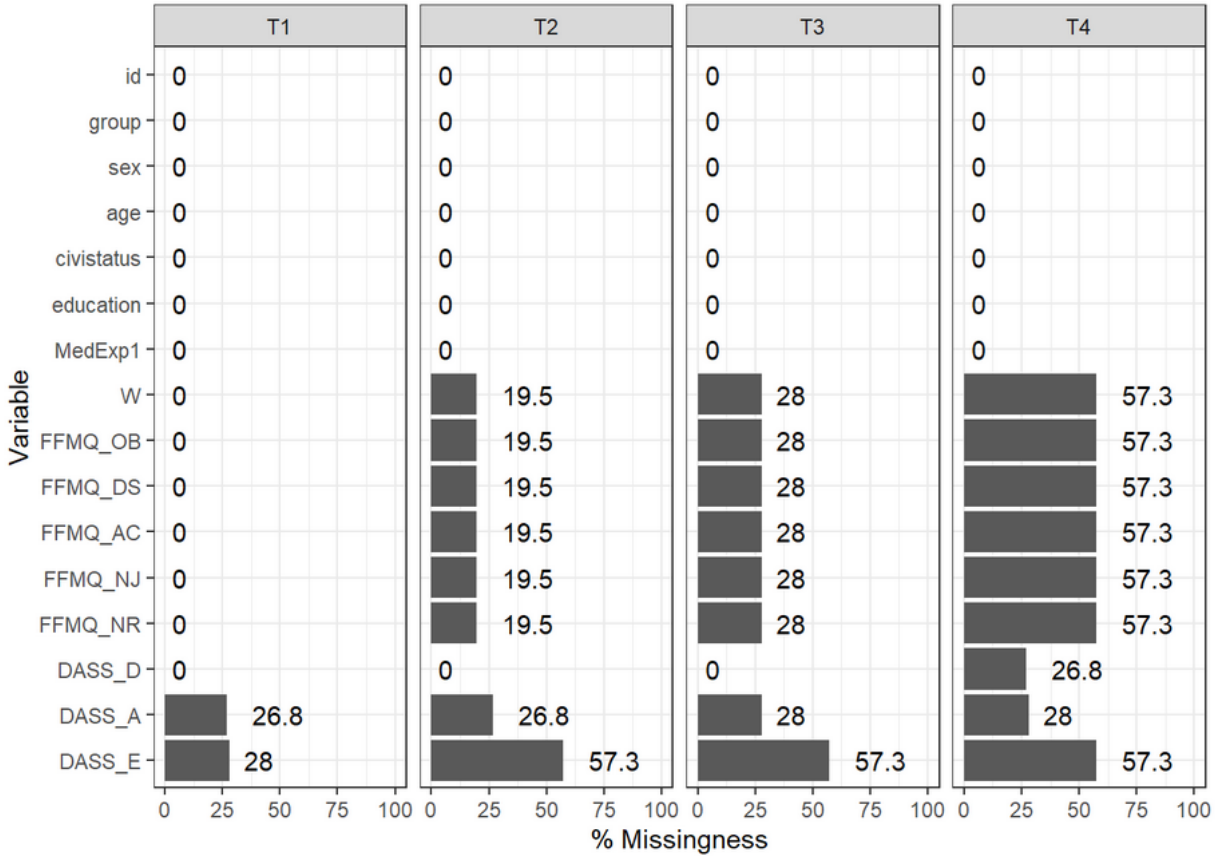


Missingness Assessment

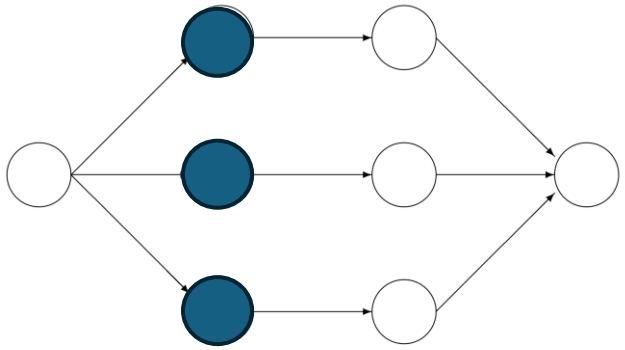


- Presence of dropout causing missingness (mean = 22,5%)
- Small differences between groups



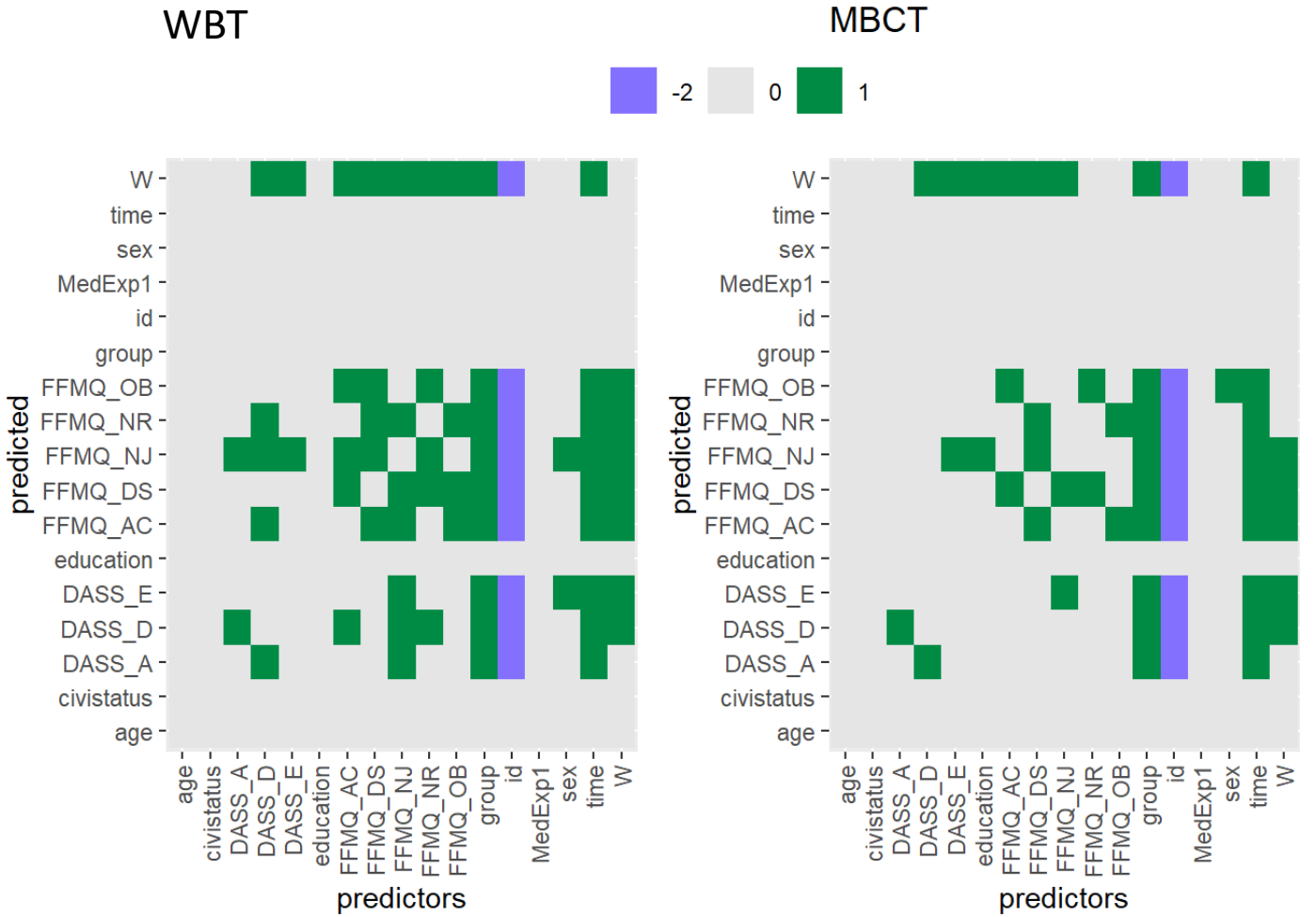


Building the imputation model

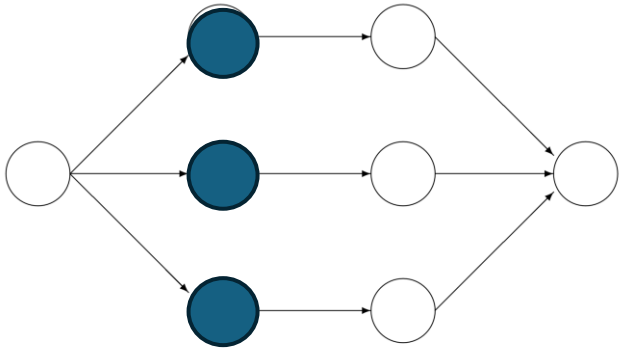


Predictor matrix

- Theoretical criteria:
  - Remove irrelevant variables (e.g., some demographics, raw ítems)
  - Group and Time (level-1), ID (level-2)
- Empirical criteria (r >.3, level 1)



Building the imputation model



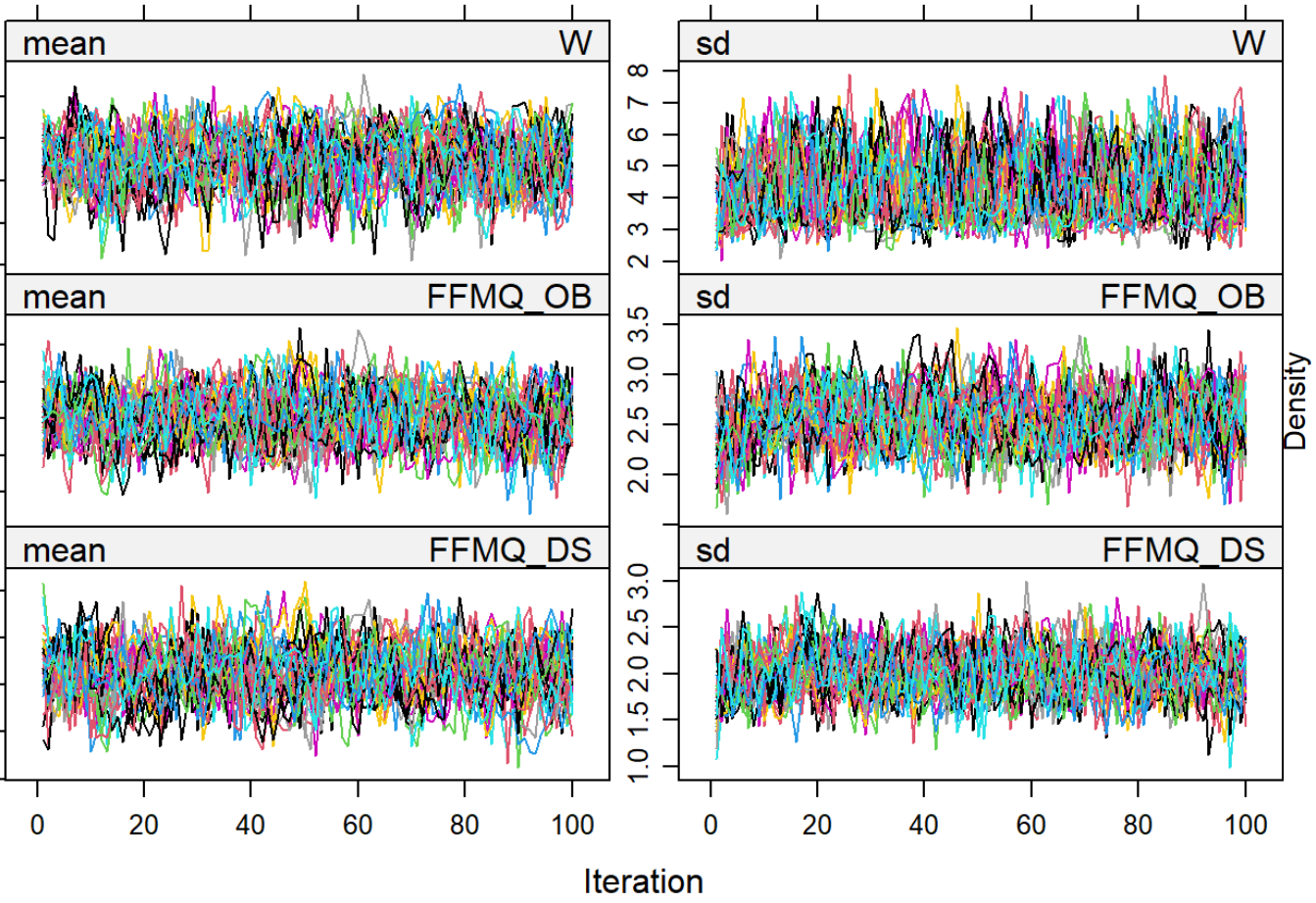
- Estimation methods
- 2-level generalist method: *2l.pmm*
  - Predictive Mean Matching adapted for 2-level data structures
  - Other methods available (*2l.pan*, *2l.jomo*,…)
- Run the model (one per group)
  - Samples: 23 (mean % of sample's missing values)
  - Iterations: 100 (increase if available)

```
impmeth <- rep("2l.pmm",ncol(data_long))
names(impmeth) <- colnames(data_long)
impmeth[c("id", "time","group","sex", "age",
          "civistatus","education", "MedExp1")] <- ""
```

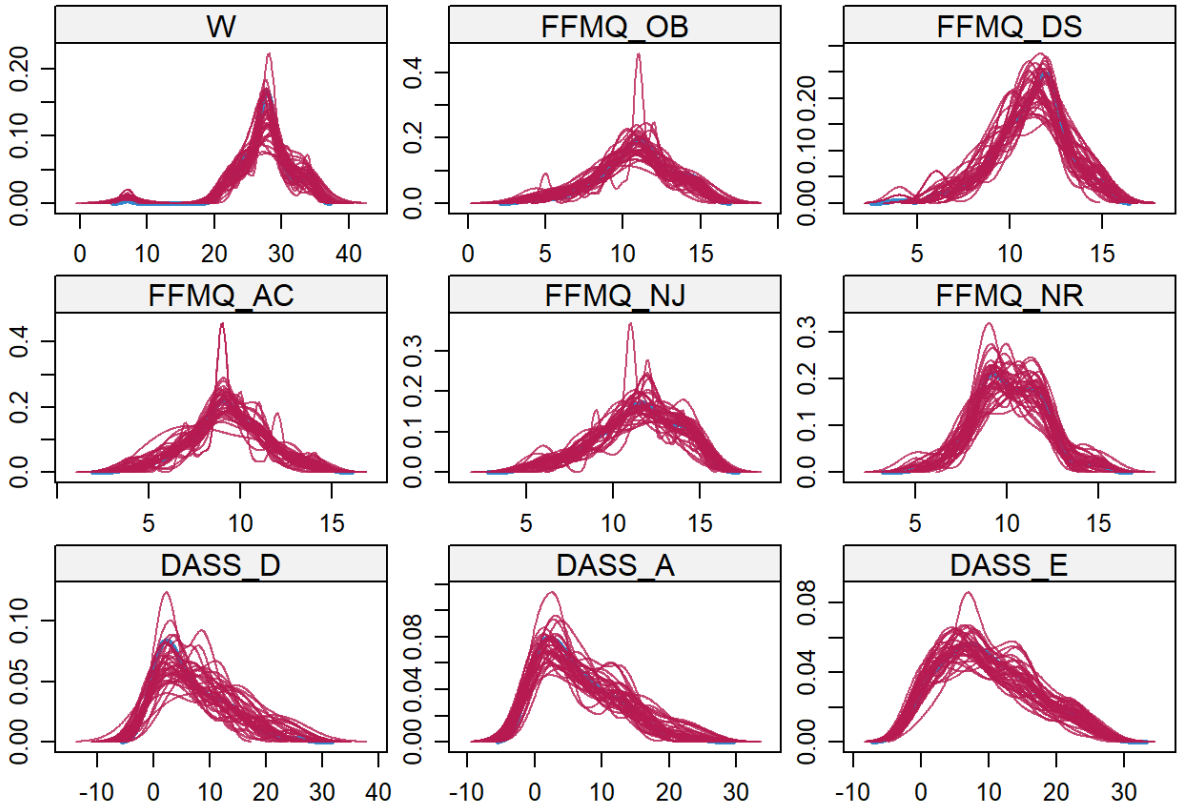
```
samples <- 23 # the (truncated) mean of missing data
maxit <- 100 # adjust according to assumable computation time. Start low and increase

impEBC <- mice::mice(dataEBC, predictorMatrix = predEBC, m = samples,
                     maxit = maxit, meth = impmeth, seed = 500)

impMBCT <- mice::mice(dataMBCT, predictorMatrix = predMBCT, m = samples,
                      maxit = maxit, meth = impmeth, seed = 500)
```

Estimating the model

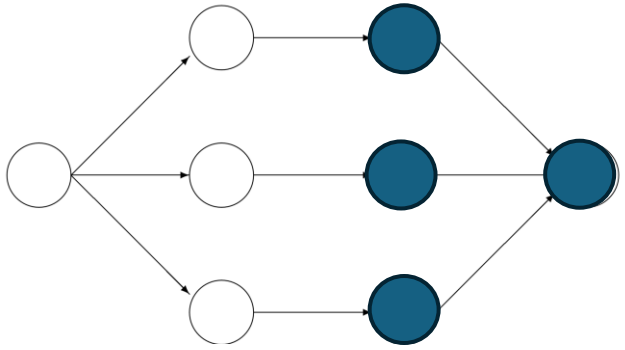
- Chains with “caterpillar” shape



- Densities similar to empirical data (unless suspecting MNAR)



Assessment of results

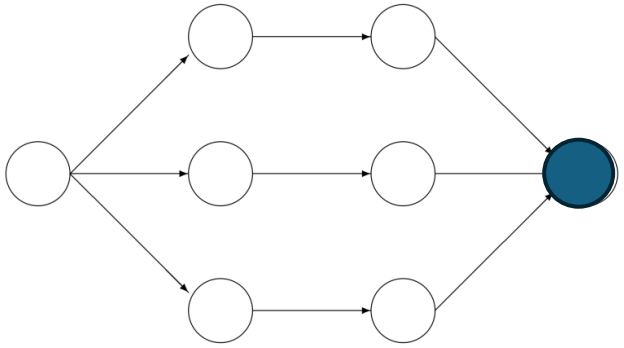


- Estimating ANOVAs as LMMs (one per sample)
- Pooling them
- Obtaining statistics (F, post hocs, EMMs)

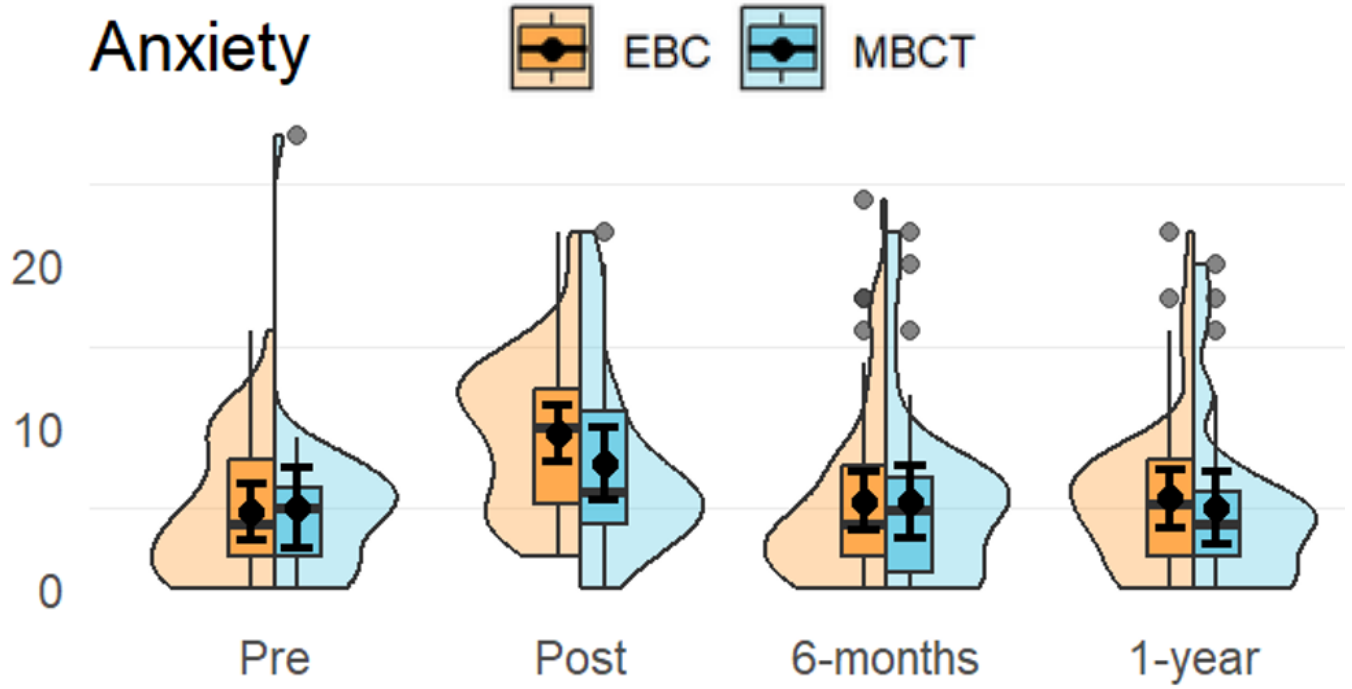
```
# run linear mixed models one per dataset
modelFitW <- with(impfull,lmer("W ~ group*time + (1|id)"))

# pool them:
pooledW <- pool(modelFitW)
pooledW

# prepare for display:
CoefTestsW <- as.data.frame(summary(pooledW))
CoefTestsW
```

...1 <chr>	term <fct>	estimate <dbl>	std.error <dbl>	statistic <dbl>	df <dbl>	p.value <dbl>
W	(Intercept)	25.869300912	1.2963731	19.955136150	315.9871	6.676820e-58
W	group	0.279635258	0.8584522	0.325743538	315.9871	7.448340e-01
W	timeT2	3.037689970	1.5195818	1.999030292	247.9792	4.669648e-02
W	timeT3	1.245141120	1.5448480	0.805995880	228.2858	4.210840e-01
W	timeT4	0.606478506	1.7882846	0.339139818	112.1775	7.351385e-01
W	group:timeT2	-0.504559271	1.0185065	-0.495391335	233.4544	6.207899e-01
W	group:timeT3	0.423551889	1.0267067	0.412534456	224.0767	6.803422e-01
W	group:timeT4	0.540026053	1.1392533	0.474017531	132.9366	6.362650e-01

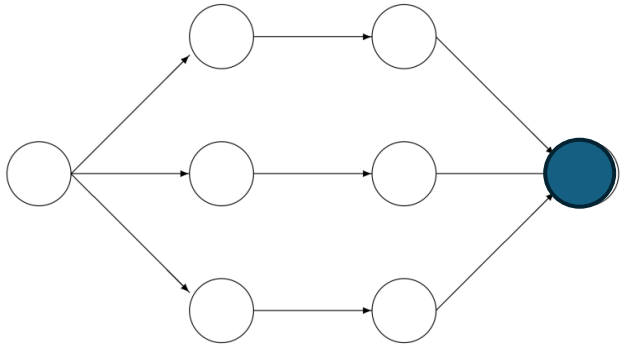
Assessments of results



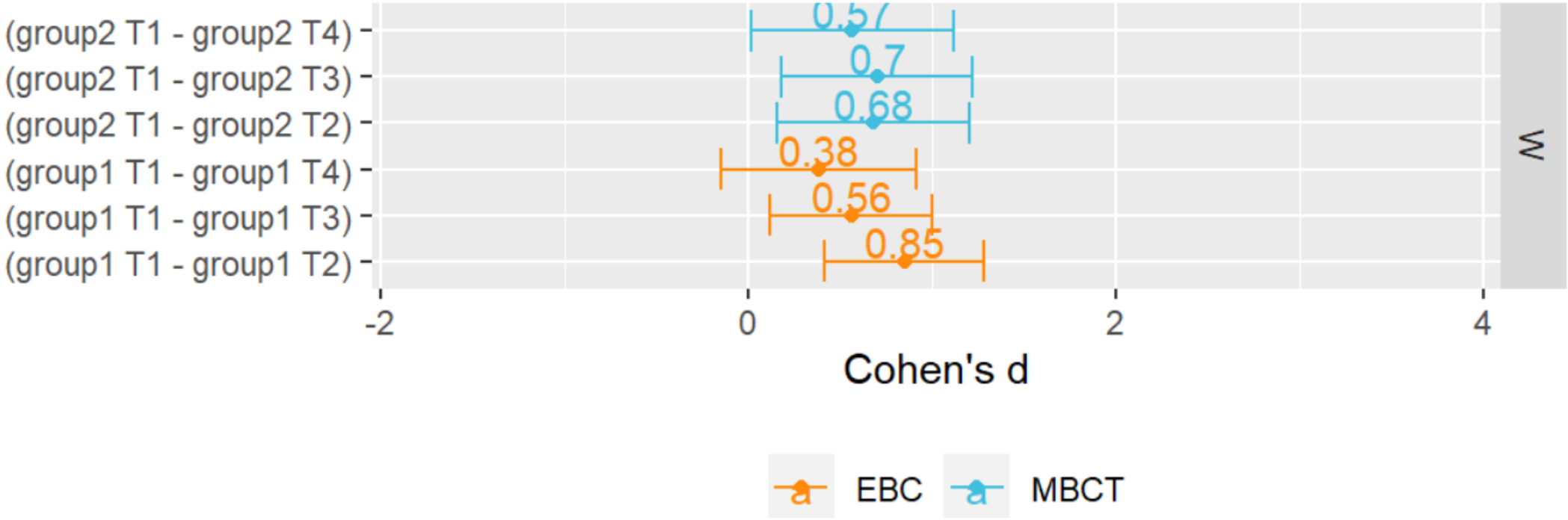
- Plotting the results
  - Completing the dataset (only to create density plots)
  - Estimate marginal means with CIs



Assessments of results

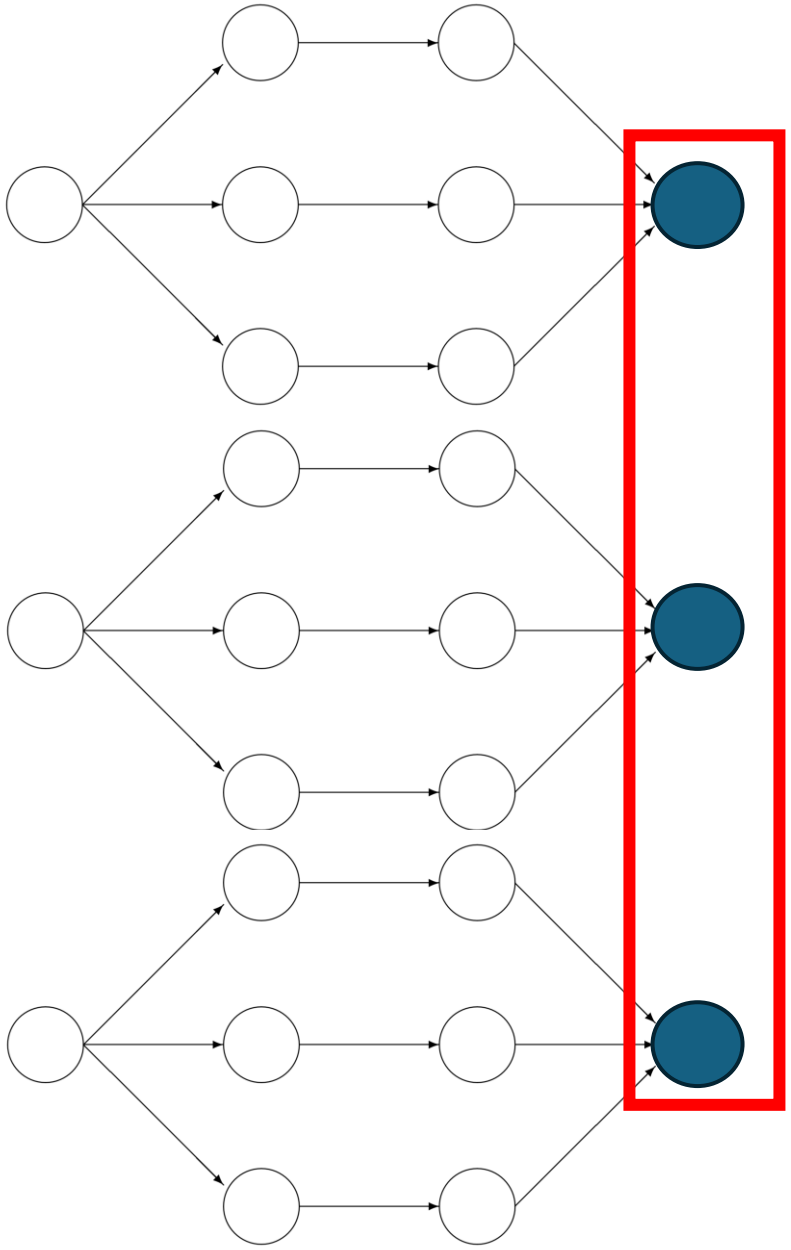


- Extra: Comparison between effect sizes
  - Obtaining each effect size from *post hocs*
  - Estimating CIs and plotting them

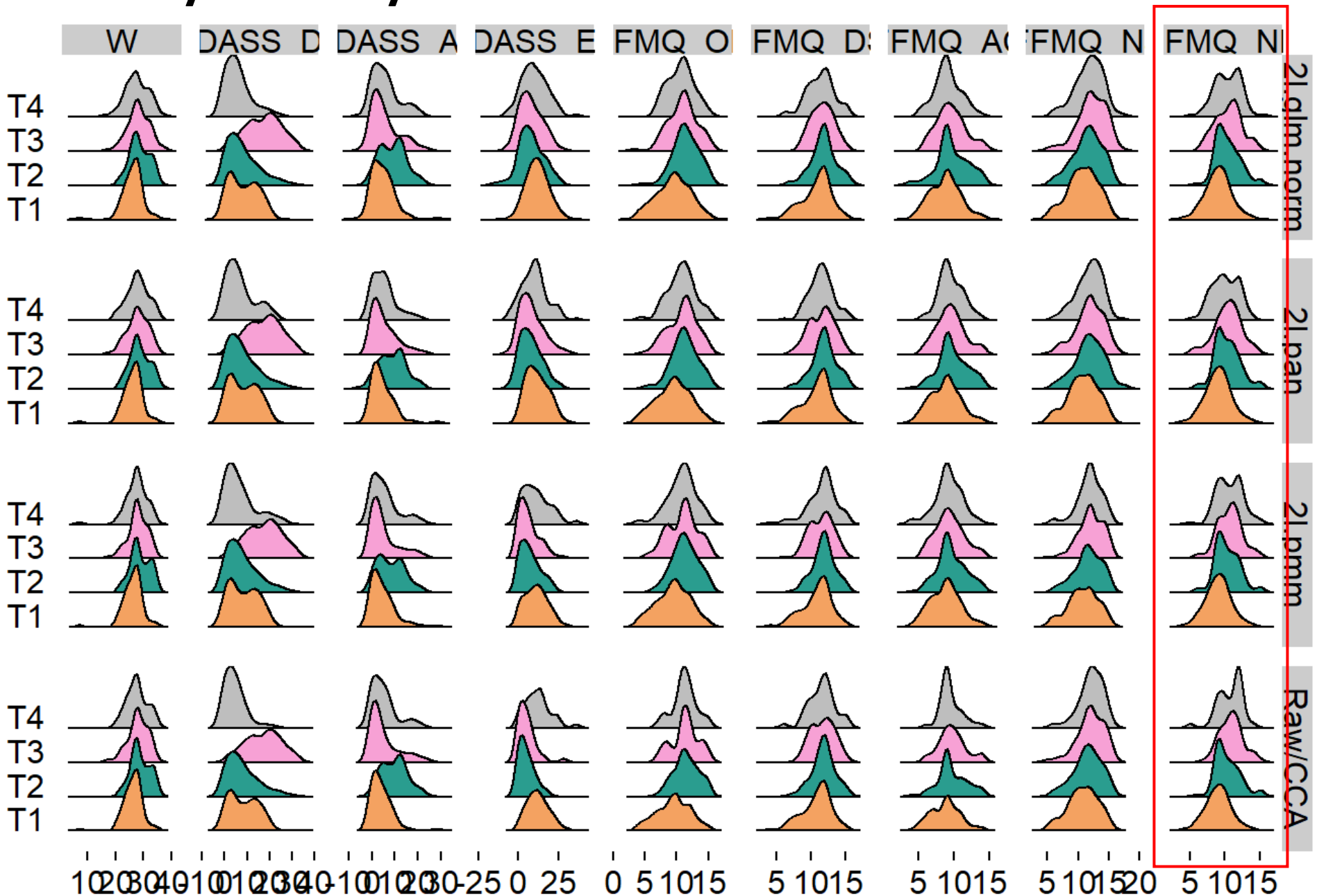


Sensitivity Analysis

- Change aspects of the model
  - Estimation method (CCA, MVNI, PMM,…
  - Predictors (>0.3 or >0.4)
  - nSamples, iterations,…
- Compare results
  - Densities, p-values, means/CIs,…
- If a difference is found
  - Method bias?
  - One is better at reproducing population distributions
    - Which one?



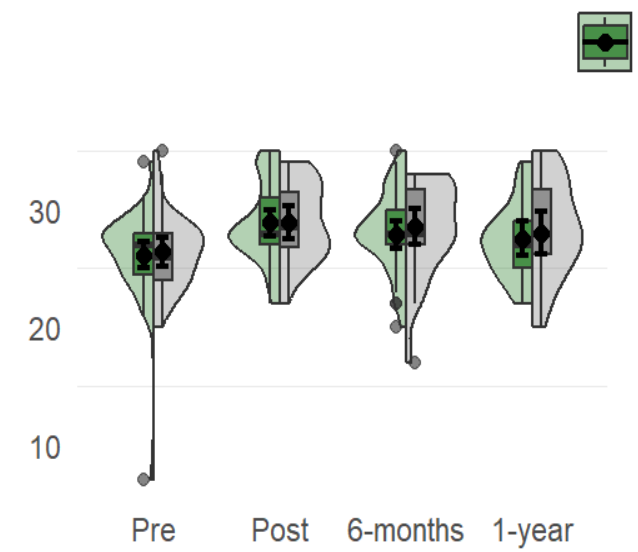
Sensitivity Analysis



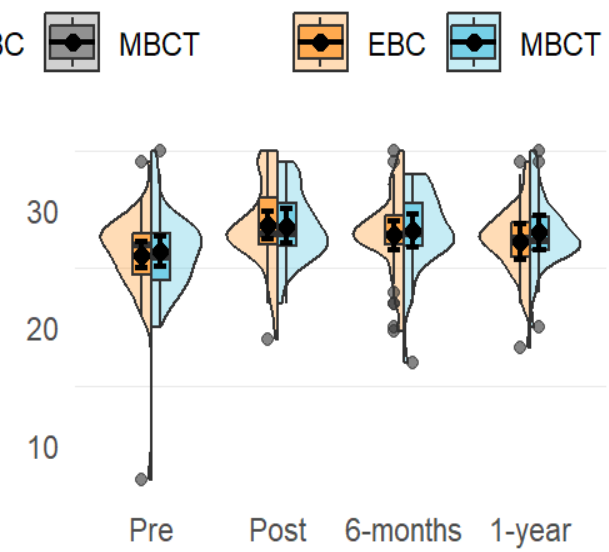
Sensitivity Analysis

- We can spot:
  - Technical issues
  - Differences between methods in producing results
  - Remember: Similar results to CCA does not imply that it is okay to use CCA (power loss is still present)

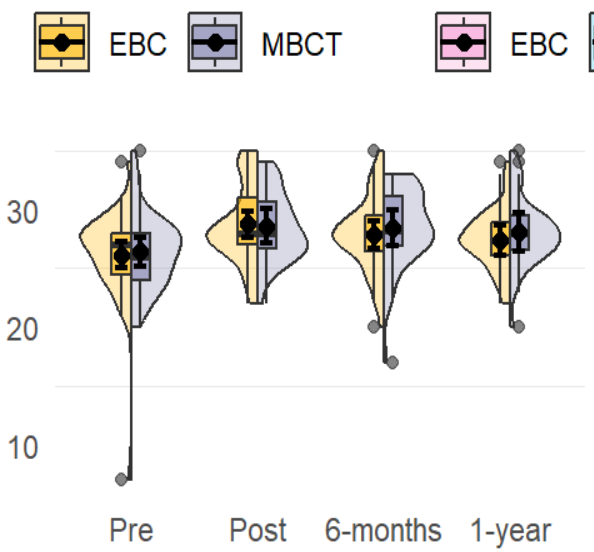
CCA



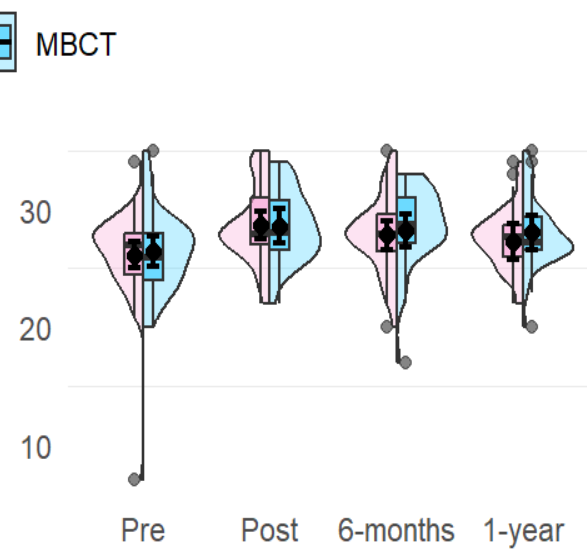
2l.PMM



2l.PMM.PAN



2l.GLM.NORM



Conclusions

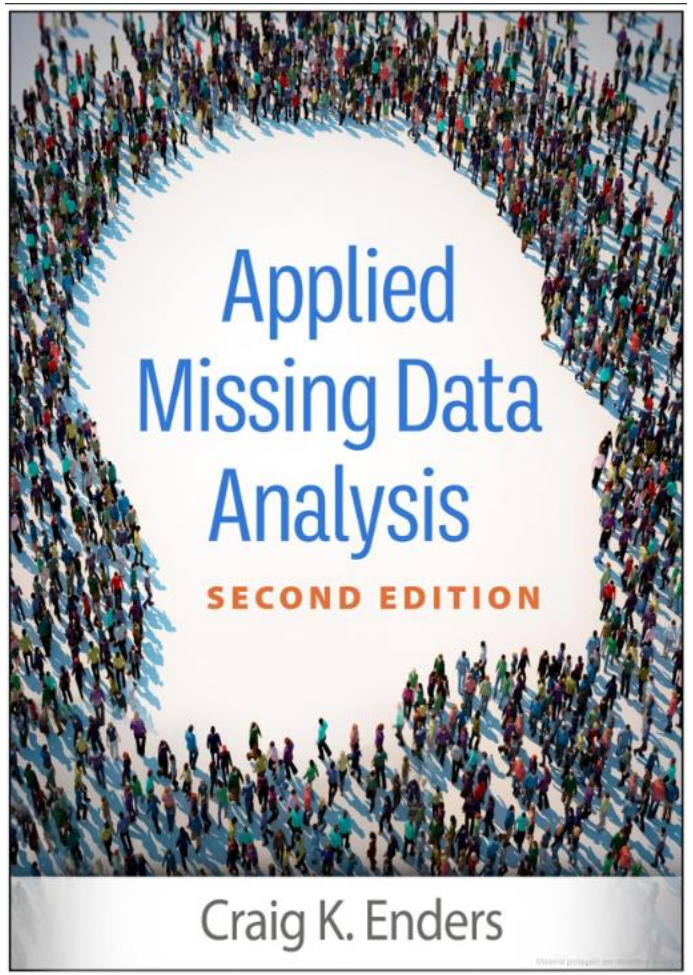
- Multiple imputation using *mice* is feasable for RCTs
- It is important to ensure rigor and care for data structure and technical specifications
- Sensitivity analysis and selection of predictors are very relevant

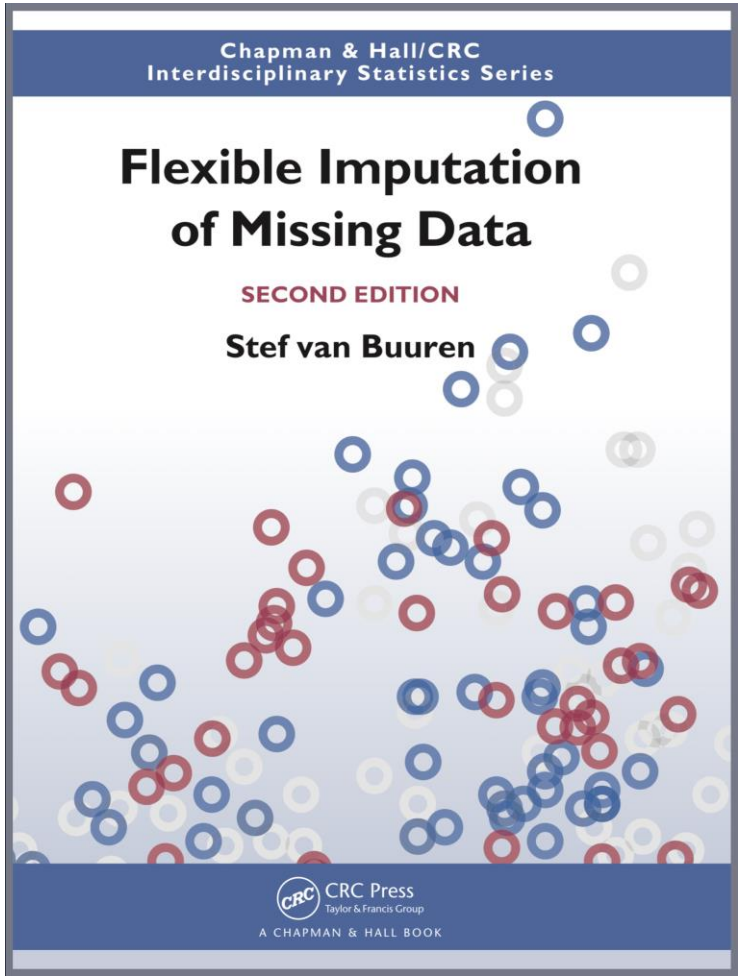
- Alternatives that may be better if data/analysis fit in:
  - *Full-Information Maximum Likelihood* (implemented in *lavaan*)
  - *Bayesian imputation* (e.g., BLIMP)

Recommended sources

- Books:
  - van Buuren (2018)
  - Enders (2022)
- Reviews:
  - Enders (2023)
  - Morel et al. (2022)
- Tutorials:
  - Austin et al. (2020)

https://stefvanbuuren.name/fimd/





Thank you!

Script and data here:

Oscar Lecuona

Ariadna Angulo-Brunet

Víctor Ciudad

Ricardo Olmos