

Methods: Mind the Gap
Webinar Series

Multiple Imputation Methods for Group-Based Interventions



Presented by:

Rebecca Andridge, Ph.D.

The Ohio State University



National Institutes of Health
Office of Disease Prevention

Outline

- Missing data in group-based intervention trials
- What is multiple imputation (MI)?
- When to use MI for group-based intervention trials
- How to use MI (correctly) for group-based intervention trials
- Software for MI in group-based trials

Missing data in group-based intervention trials



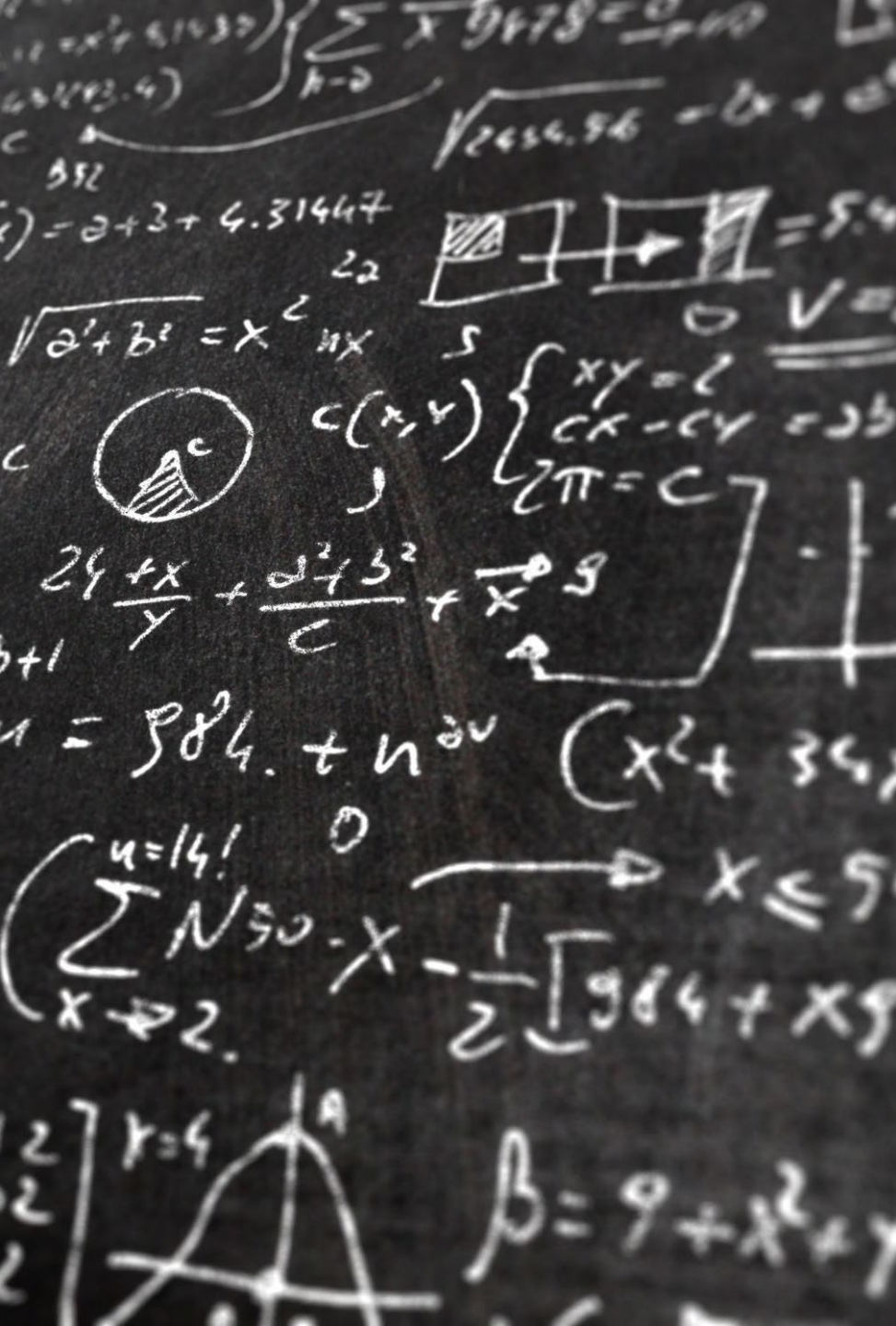
Randomized Trial Designs for Group-Based Interventions

- Cluster randomized trials (CRTs)
 - Identifiable groups rather than individuals are randomized to study conditions
- Individually randomized group treatment trials (IRGTs)
 - Individuals randomized, but intervention delivered in groups
- Stepped wedge cluster randomized trials (SW-CRTs)
 - Identifiable groups sequentially transition from control to intervention
- **Possible advantages:** reduce treatment contamination, lower costs, test intervention when individual randomization not possible, deliver intervention to everyone (SW-CRTs)

Resulting Data

- Usually, a small number of groups (clusters) with **correlated observations** within a group
 - Cluster characteristics make individuals “more similar” within group
 - Intervention delivered in a group, causing “more similar” intervention effects within the group
- Resulting intraclass correlation (ICC) is often small, but can lead to large variance inflation factors and cannot be ignored
 - Design effect = $[1 + (m - 1)\rho]$ where m = size of cluster and ρ = ICC

Ignore clustering → Variances too small → p-values too small!



Possible Analysis Methods

- **Cluster-level analyses**
 - “Remove” the correlation for analysis through summary statistics
- **Individual-level analyses**
 - Mixed effects (random effects) models – explicitly model the within-cluster correlation via random effects
 - Generalized estimating equations – treat correlation as a “nuisance” and use robust inference
- Complete data from all participants? 😊 All methods are appropriate
- Missing data from some participants? ☹️ Methods may not be appropriate except under (strong?) assumptions

Missing Data in CRTs

- Quite common!
- Two types 
- Whole clusters missing is a major challenge...
- Also have two types of variables:
 - Individual-level
 - Cluster-level variables
- I'll focus on missing *individual-level* data

Individuals with missing data

CLUSTER	TRT	X	Y	Z
1	0	90	46	0
1	0	???	???	???
1	0	5	18	1
1	0	???	35	1
2	0	9	18	0
2	1	56	???	0
2	1	77	33	1
2	1	82	21	???
2	1	41	17	0
2	1	75	25	1
3	0	17	17	1
3	0	???	???	???
3	0	???	???	???
3	0	41	23	1
4	1	94	18	1
4	1	48	25	0
4	1	???	???	???
4	1	50	???	1
4	1	30	28	1

Entire clusters with missing data

CLUSTER	TRT	X	Y	Z
1	0	90	46	0
1	0	9	12	1
1	0	5	19	1
1	0	63	35	1
2	0	9	18	0
2	1	56	25	0
2	1	77	33	1
2	1	82	21	1
2	1	41	17	0
2	1	75	25	1
3	0	???	???	???
3	0	???	???	???
3	0	???	???	???
3	0	???	???	???
4	1	94	18	1
4	1	48	25	0
4	1	71	25	0
4	1	50	24	1
4	1	30	28	1

Lingo: Analysis Strategies w/Missing Data

- **Complete case analysis** – use observations that are complete with respect to **all variables**
- **Available case analysis** – use observations that are complete with respect to **variables used in the analysis**
- **Imputation** – fill in missing values to create completed data set(s)

Complete Case Analysis

CLUSTER	TRT	X	Y	Z
1	0	90	46	0
1	0	????	????	????
1	0	5	18	1
1	0	????	35	1
2	0	9	18	0
2	1	56	????	0
2	1	77	33	1
2	1	82	21	????
2	1	41	17	0
2	1	75	25	1
3	0	17	17	1
3	0	????	????	????
3	0	????	????	????
3	0	41	23	1
4	1	94	18	1
4	1	48	25	0
4	1	????	????	????
4	1	50	????	1
4	1	30	28	1

Available Case Analysis



CLUSTER	TRT	X	Y	Z
1	0	90	46	0
1	0	????	????	????
1	0	5	18	1
1	0	????	35	1
2	0	9	18	0
2	1	56	????	0
2	1	77	33	1
2	1	82	21	????
2	1	41	17	0
2	1	75	25	1
3	0	17	17	1
3	0	????	????	????
3	0	????	????	????
3	0	41	23	1
4	1	94	18	1
4	1	48	25	0
4	1	????	????	????
4	1	50	????	1
4	1	30	28	1



Lingo: Missing Data Taxonomy

MCAR: Missing Completely At Random

- Probability a variable has missing values **does not depend on any variable** (including itself)
- Subjects with missing data are **indistinguishable** from those with complete data



MAR: Missing At Random

- Probability a variable has missing values **depends on observed variables' values**
- Subjects with missing data are different from those with complete data, **but we can see how they are different**



MNAR: Not Missing At Random

- Probability a variable has missing values **depends on values of unobserved variables**
- Subjects with missing data are different from those with complete data, **but we do not observe the characteristic(s) that make them different**

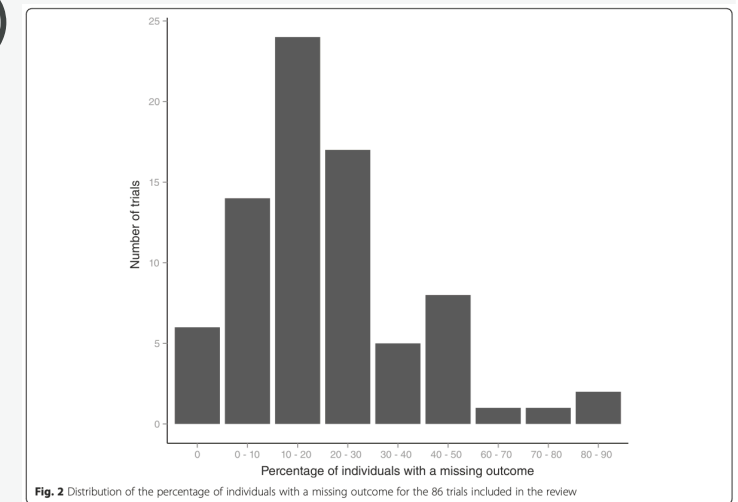


State of Missing Data Handling in CRTs

- Fiero et al. systematic review¹ – 86 CRTs
 - Nearly one-third (31%) had entire clusters missing (!!)
 - Majority (93%) reported at least some individual-level missing data
 - Median = 19% of subjects missing the outcome
 - Complete case analysis most used method (55%)
 - Only 2% of trials used multiple imputation (MI)
 - Neither of those trials used MI **correctly**

¹Fiero, M. H., Huang, S., Oren, E., & Bell, M. L. (2016). Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*, 17, 72. <https://doi.org/10.1186/s13063-016-1201-z>

Figure 2 from Fiero (2016)



What is multiple imputation (MI)?



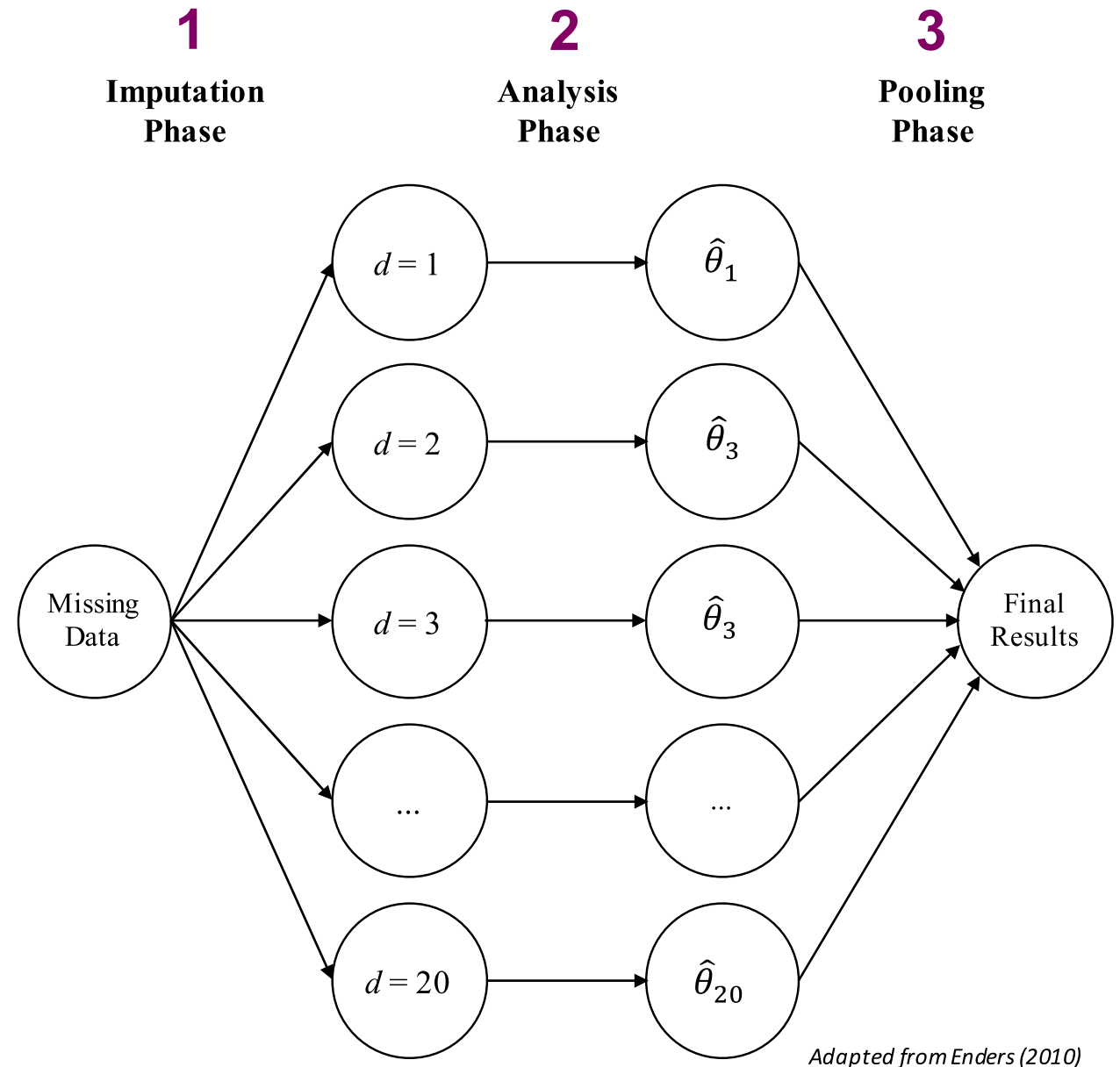
Multiple Imputation is...

- ... **not** trying to exactly replicate unobserved data values
- ... an attempt to use all available information in your data
- ... most useful when you have good predictors of missing data values
- ... flexible/adaptable for different data types
- ... able to be used for missing outcome and/or covariate data
- ... usually performed under an **MAR assumption**
- ... **not** necessarily the “best” solution for a randomized trial

Multiple Imputation

1. Create D completed data sets ($D \geq 2$) using **appropriate** imputation procedure
 - Each set a draw from the predictive distribution of missing values
2. Analyze each completed data set **separately**
3. Combine results in a (relatively) straightforward way
 - “Rubin’s Rules” (Rubin, 1987)

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.



MI Inference for a Scalar Estimand

- $\hat{\theta}_d$ = estimate of θ from d th completed dataset (e.g., trt effect)
- W_d = estimate of variance of θ from d th completed dataset
- MI estimate of θ : $\theta_D = \frac{1}{D} \sum_{d=1}^D \hat{\theta}_d$ ← Average of the D point estimates
- MI estimate of variance: $T_D = W_D + \left(1 + \frac{1}{D}\right) B_D$
 - W_D = Within-imputation variance = $\frac{1}{D} \sum_{d=1}^D W_d$ ← Average of the D variance estimates
 - B_D = Between-imputation variance = $\frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \theta_D)^2$ ← Variance of the D point estimates
- Inference: $\bar{\theta}_D \sim t_\nu$ where $\nu = (D - 1) \left[1 + \frac{W_D}{\left(1 + \frac{1}{D}\right) B_D} \right]^2$

Note: there is an alternative formula for ν when complete data DF are small – like in CRTs

Example of MI (D=3)

Y	Z
15	11
25	13
?	10
13	8
13	7
?	9
16	10
20	13
21	10
?	8



$d = 1$	
Y	Z
15	11
25	13
20.1	10
13	8
13	7
13.6	9
16	10
20	13
21	10
12.6	8

$\bar{y} = 16.93$
 $\hat{V}(\bar{y}) = 1.85$

$d = 2$	
Y	Z
15	11
25	13
13.1	10
13	8
13	7
14.3	9
16	10
20	13
21	10
9.9	8

$\bar{y}$
 $\hat{V}(\bar{y}) = 2.11$

$d = 3$	
Y	Z
15	11
25	13
19.3	10
13	8
13	7
11.1	9
16	10
20	13
21	10
14.4	8

$\bar{y} = 16.78$
 $\hat{V}(\bar{y}) = 1.92$

- Estimate of mean of $Y = (16.93 + 16.03 + 16.78)/3 = 16.58$
- Between-imputation variance = $B = \text{variance of } \bar{y}'\text{s} = 0.23$
- Within-imputation variance = $W = (1.85 + 2.11 + 1.92)/3 = 1.96$
- Total variance = $T = W + \left(1 + \frac{1}{D}\right) B = 1.96 + \left(1 + \frac{1}{3}\right) 0.23 = 2.27$

Notes About MI

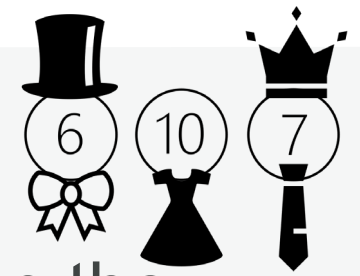
Imputations will (must!) vary across MI data sets

Imputations must be **proper** for correct inference

Imputations can come from **implicit** or **explicit** models

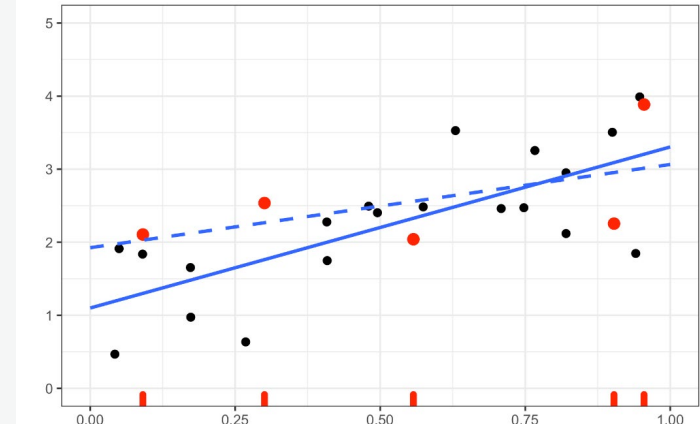
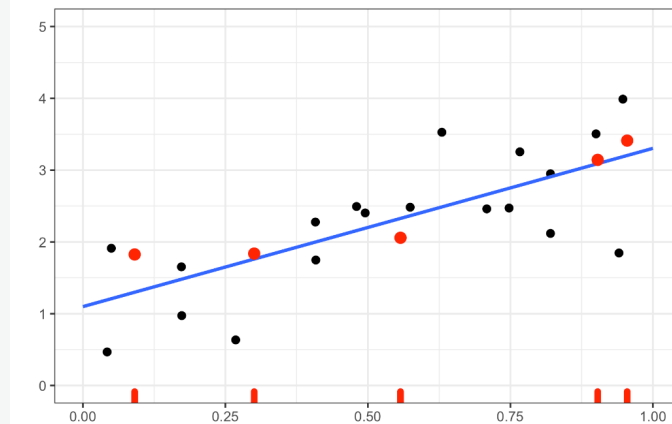
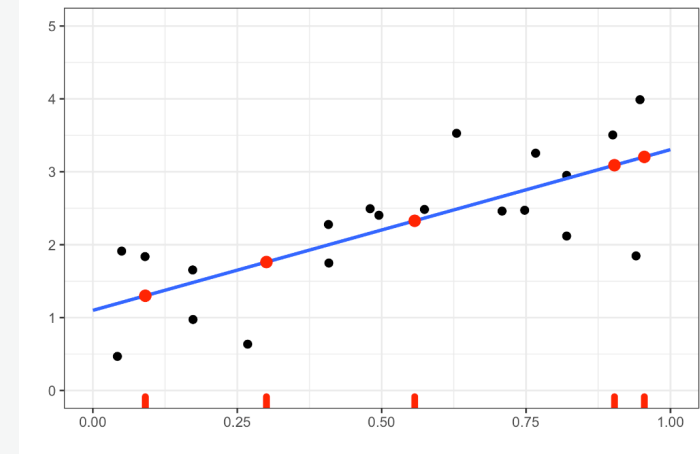
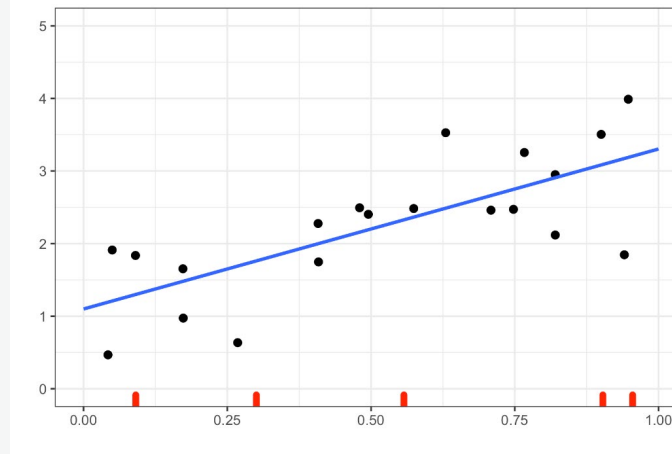
Imputation model must be **congenial** to the analysis model





“Proper” MI – What’s that?

- Imputations must *propagate the uncertainty* inherent in the models being used



Congeniality – What's that?

- Two models involved:
 - Imputer's model: model(s) for missing variable(s)
 - Analyst's model: answering the scientific question
- A **congenial** imputation model is *consistent with or compatible with* the analysis model
 - Can the analyst's model be derived from the imputer's?
(*Is the analyst's model a "subset" of the imputer's?*)
 - Yes → **congenial** ✓
 - No → **uncongenial** ⊗

Uncongeniality in Group-Based Trials

- Sources of possible uncongeniality:
 - Omitting covariates
 - Not including treatment indicator
 - Not including outcome(s) when imputing covariates
 - Leaving out interactions (if they really exist)
 - **Ignoring clustering**



Two Main Types of MI

- **Joint Model (JM)**
 - Specify multivariate distribution for all variables with missingness
 - Simultaneously draw imputations from this multivariate distribution
- **Fully Conditional Specification (FCS)**, a.k.a., Multiple Imputation by Chained Equations (MICE) or Sequential Regression Imputation (SRI)
 - Specify separate conditional distributions for each variable with missingness
 - Impute sequentially from each distribution
- Equivalent with single-level multivariate normal data (but not other types)

When to use MI for group-based intervention trials



Key Fact: Complete Case Sometimes Okay

- Complete case analysis yields unbiased estimates of:
 - *All estimates* when data are MCAR (missing Y 's or X 's)
 - *Regression coefficients* with missing Y 's where missingness depends on X
- So, if missingness in Y depends only on a baseline covariate X or treatment assignment T (both fully observed), then a Complete Case analysis yields unbiased estimates in the model:

$$E[Y] = \beta_0 + \beta_T T + \beta_X X$$

- MI approximates Complete Case if you use it in this case

So, When Would We Want to Use MI?

- If you believe MAR (not MCAR) and...
- ... you *don't want to condition on X* in your analysis model
 - Can use X to impute missing Y , then run model without X
- ... you want to “use all randomized participants in the analysis”
 - Some have argued that this is necessary for analysis to be Intent to Treat (ITT)
- ... you have some missingness in X (less likely for baseline covariates)
- In CRTs/IRGTs/SW-CRTs, analysis model is often specified *a priori*
 - If regression model for the outcome does not include covariates related to missingness (and to Y), then complete case estimates likely to be biased
 - MI a sensible approach

How to use MI (correctly) for group-based intervention trials



UNCONGENIAL Strategies for MI in CRTs

Imputation Model Strategy	Comments
Ignore clustering	Variance estimates too small – worse with larger ICCs P-values too small Source: Taljaard, M., Donner, A. and Klar, N. (2008). Imputation strategies for missing continuous outcomes in cluster randomized trials. <i>Biometrical Journal</i> 50, 329–345.
Fixed effects for clusters (dummy coding/indicator variables for cluster) (remove intervention group indicator)	Variance estimates too large – worse with smaller ICCs P-values too big Source: Andridge, R.R. (2011) Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. <i>Biometrical Journal</i> 53, 57–74.



How “Bad” Are These MI Methods?

- Post-test only CRT with $l = 1, 2$ treatment groups
- Each treatment group has $j = 1, \dots, k$ clusters
- Each cluster has $i = 1, \dots, m$ individuals per cluster
- True model:

$$y_{ijl} = \mu_l + b_{jl} + e_{ijl}$$

$$b_{jl} \sim N(0, \rho\sigma^2) \quad \text{and} \quad e_{ijl} \sim N(0, (1 - \rho)\sigma^2), \quad b_{jl} \perp e_{ijl}$$

$$\rho = \text{ICC}$$

$$\text{Design effect} = [1 + (m - 1)\rho]$$

- MCAR missing data
- Percentage of units with observed data in each cluster = π



How “Bad” Are These MI Methods?

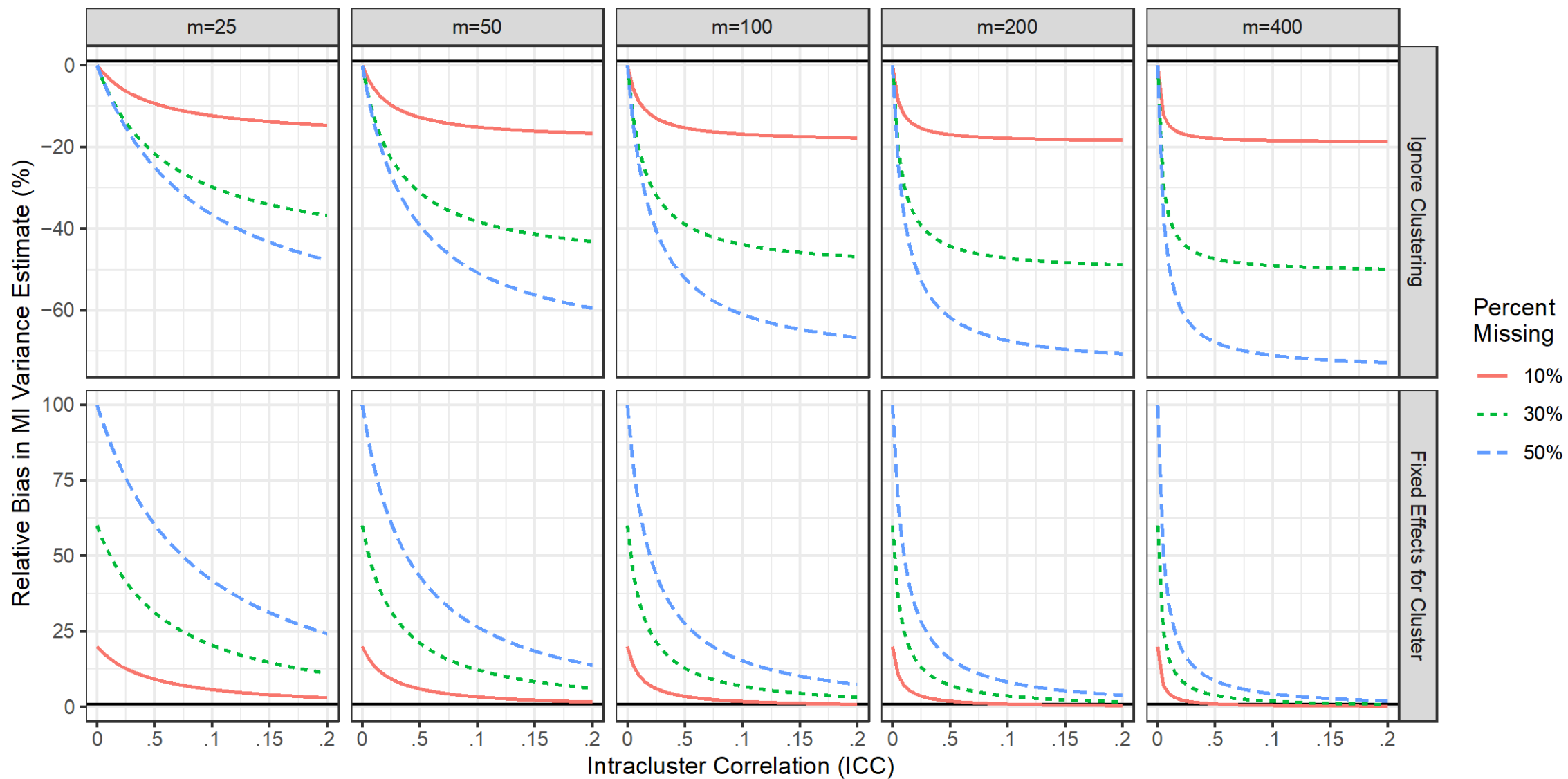
- Impute **ignoring** clusters:

- Bias of MI variance estimate $= \frac{\sigma^2 \rho}{km\pi - 1} (m\pi - 2)(\pi^2 - 1)$
- Bias always **negative** (*underestimate variance*)
- Bias **worse for larger ICCs (ρ)**

- Impute with **fixed effects** for clusters:

- Bias of MI variance estimate $= \frac{2(1-\pi)(1-\rho)\sigma^2}{km\pi}$
- Bias always **positive** (*overestimate variance*)
- Bias **worse for smaller ICCs (ρ)**





CONGENIAL Strategies for MI in CRTs

Imputation Model Strategy	Comments
Mixed model with random intercept for cluster (a specific type of “Multilevel MI”)	PROS: • Congenial to mixed model for analysis! 😊 • Software: Available in R and BLIMP CONS: • Might have a small number of clusters with which to estimate random effect variance • Software: Not available in SAS, Stata, SPSS
Predictive Mean Matching adapted for CRTs New method developed in: Bailey, B.E., Andridge, R. & Shoben, A.B. (2020) Multiple imputation by predictive mean matching in cluster-randomized trials. <i>BMC Med Res Methodol</i> 20, 72.	PROS: • Relaxes model assumptions • Used a model only to identify close “donors,” not for generating imputations CONS: • Software: Not available (user-written only)

Illustrative CRT: Health Education for Hypertension

- CRT in South India in 2015-2016 (data publicly available¹)
 - Intervention = Community Health Worker (CHW)-led education program
 - Control = Usual care
- Clusters = villages
 - Intervention = 5 villages (median size = 134, range 97-154)
 - Control = 10 villages (median size = 109, range 93-139)
 - Total sample size = 1,734 (636 intervention, 1,098 control)
- Primary Outcome = Control of hypertension
 - Measured ~2 months post-intervention

¹ Gamage DG, Riddell MA, Joshi R, Thankappan KR, Chow CK, Oldenburg B, et al. (2020) Effectiveness of a scalable group-based education and monitoring program, delivered by health workers, to improve control of hypertension in rural India: A cluster randomised controlled trial. *PLoS Med* 17(1): e1002997.

Missing Data for Follow-up Outcome

- Outcome = Change in SBP, follow-up minus baseline
- Missing SBP data at follow-up for ~15% (264/1,734)
 - Treatment: 28% (178/636)
 - Control: 8% (86/1,098)  Differential – cause for concern?
- Evidence *against* MCAR:

BASELINE Variable	Observed FU SBP (n=1470)	Missing FU SBP (n=264)	p-value
Baseline SBP, mean (SD)	137.8 (22.3)	144.5 (22.3)	.001
% Controlled HTN	51.4%	26.9%	<.0005
% Using BP meds	42.6%	23.1%	<.0005
% Never had routine MD check-up	16.6%	36.9%	<.0005
% Male	40.2%	50.8%	.0001

Predictors of Outcome (among CC)

- Importantly, many of these baseline variables are associated with the outcome **among complete cases**

BASELINE Variable	Association with Change in SBP		p-value
Baseline SBP	correlation = -0.50		<.0005
% Controlled HTN	Not Controlled Controlled	-12.0 (22.1) 3.6 (16.1)	<.0005
% Using BP meds	No BP Meds Yes BP Meds	-6.1 (19.8) -1.1 (21.8)	<.0005
% Never had routine MD check-up	Had checkup No checkup	-3.0 (20.6) -8.6 (21.1)	.014
% Male	Female Male	-3.2 (19.8) -5.3 (22.3)	.077

Data shown are mean (SD) by level of predictor

Complete Case Analysis

- Outcome = Difference in SBP, follow-up minus baseline
 - Intervention group mean (SD): -8.1 (19.9)
 - Control group mean (SD): -2.1 (20.9)
- Results from mixed effects models with random intercept for cluster

Model	Estimated Mean (SE)	Treatment Effect 95% CI	p-value	ICC
Unadjusted	-6.5 (2.7)	(-12.4, -.57)	.034	.048
Adjusted for Baseline SBP	-5.7 (1.8)	(-9.5, -1.8)	.008	.023
Adjusted for variables on previous slide	-5.7 (1.7)	(-9.3, -2.0)	.005	.019

Multiple Imputation Analyses

- Impute follow-up SBP (then calculate the change score)
- Include all baseline variables on previous slide, plus age and BMI
 - Note: I did not do a comprehensive search to find the best variables!
- D = 100 imputations using FCS
- Results from unadjusted model

Imputation Strategy	Estimated Mean (SE)	Treatment Effect 95% CI	p-value	ICC
Random intercept for clusters	-7.0 (2.8)	(-13.1, -0.85)	.029	.049
Ignore clustering	-6.8 (2.5)	(-12.3, -1.3)	.020	.039 ↓
Fixed effects for clusters	-7.2 (3.0)	(-13.8, -0.68)	.017	.057 ↑

Warning: Effect Modification

- Suppose a baseline covariate is a treatment effect modifier
 - → “truth” is an interaction: **treatment group x covariate** predicting outcome
- Imputation model should **include interaction**
- If not, overall treatment effect estimate will be biased *if missingness depends on X*
- **General recommendation:** Impute separately by treatment arm¹ but...
- ... this may be challenging for group-based trials
 - Small number of clusters per treatment arm

¹Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P., & Lee, K. J. (2018). Should multiple imputation be the method of choice for handling missing data in randomized trials? *Statistical Methods in Medical Research*, 27(9), 2610–2626.



Software for MI in group-based trials



Software for Random Effects MI

- R (<https://cran.r-project.org/>)
 - Free software
 - Available for multiple operating systems (Mac, Windows, Linux)
 - `mice` package – many variations of multilevel MI
 - E.g., `mice.impute.2l.pan` (JM), `mice.impute.2l.lmer` (FCS)
 - <https://stefvanbuuren.name/fimd/ch-multilevel.html> (Online version of textbook¹)
 - `mdmb` package – FCS imputation
 - `jomo` package – JM imputation
- Not as “user-friendly” as SAS, Stata, etc. – but help files are (usually) helpful
- Recommend use of R Studio (<https://www.rstudio.com/>)

Software for Random Effects MI

- **BLIMP** (<https://www.appliedmissingdata.com/blimp>)
 - Free software
 - Available for multiple operating systems (Mac, Windows, Linux)
 - FCS imputation
 - Designed for psychological/behavioral applications
 - Analysis models more complex than typical group-based intervention trials
 - Capable of more complex multilevel imputation models (e.g., random slope models)
 - Second edition of textbook¹ coming soon with myriad examples
 - Note: Enders' textbook is a *very user-friendly* book! Great intro to missing data and imputation.

¹Enders, C.K. (2022). *Applied missing data analysis, 2nd Edition*. Guilford Press.

Software for Random Effects MI

- **Mplus** (<http://www.statmodel.com/>)
 - Not free (but, comes with excellent tech support!)
 - Available for multiple operating systems (Mac, Windows, Linux)
 - JM or FCS imputation
- **REALCOM-Impute**
(<http://www.bristol.ac.uk/cmm/software/realcom/>)
 - Not free (except to UK academics), need MATLAB to do imputation
 - JM imputation

Questions?

