Methods for Handling Missing Data in Cluster Randomized Trials

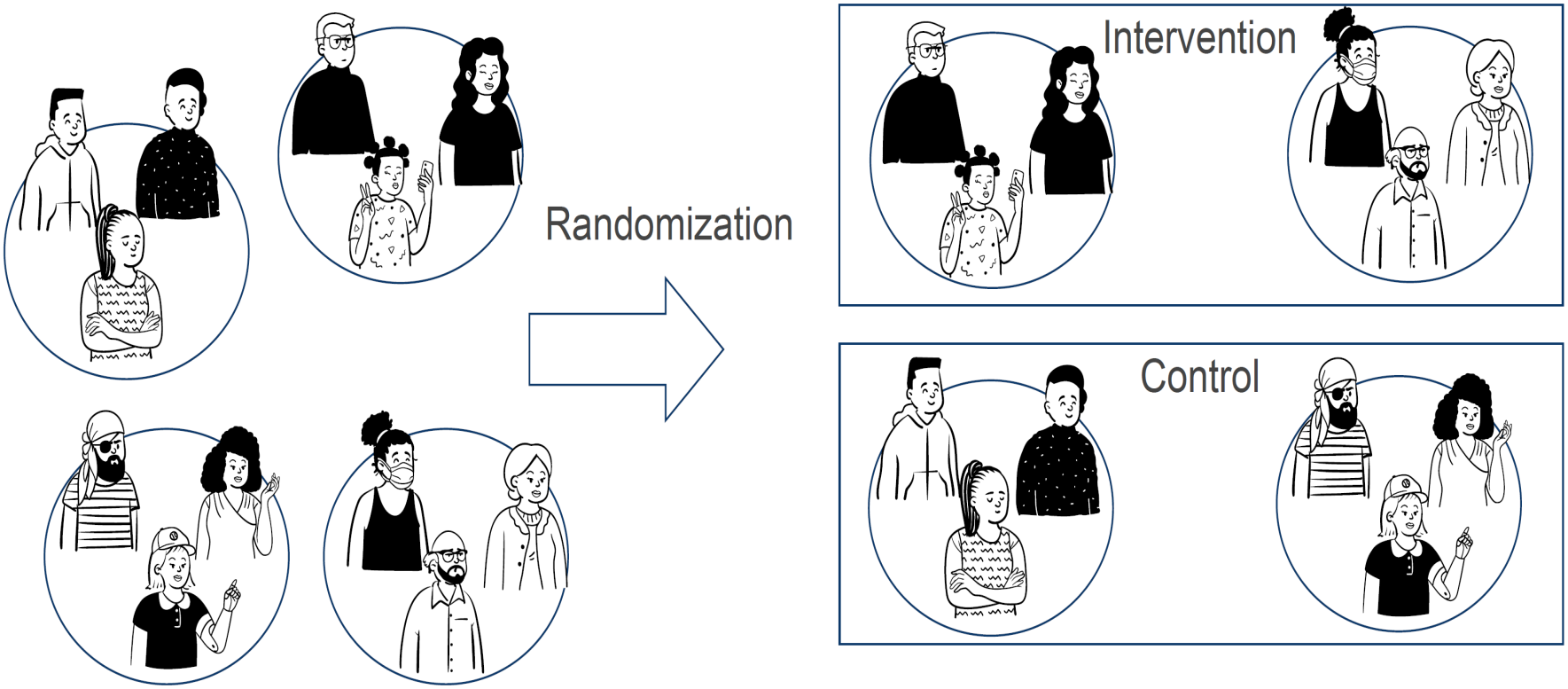
Rui Wang, Ph.D.

January 5, 2024

DEPARTMENT OF POPULATION MEDICINE
HARVARD MEDICAL SCHOOL
Harvard Pilgrim Health Care Institute
DIVISION OF BIOSTATISTICS

CRT: Randomized in groups or clusters

Cluster randomized trials are experiments in which intact social units or clusters of individuals rather than independent individuals are randomly allocated to intervention groups.



Logistic convenience and acceptability; avoid contamination

Source of the Figure: Moyer, Jonathan, "The Perils and Pitfalls of Complex Clustering in Pragmatic Trials", Available at

https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/GR-Slides-11-03-23.pdf

Data structure in CRTs

Cluster	Unit	Covariates		Treatment	Outcome
		Individual-level	Cluster-level	Indicator (A)	
1	1	X_{11}	Z_1	1	Y_{11}
1	2	X_{12}	Z_1	1	Y_{12}
...	...	...	...	...	...
1	n_1	X_{1n_1}	Z_1	1	Y_{1n_1}
2	1	X_{21}	Z_2	0	Y_{21}
2	2	X_{22}	Z_2	0	Y_{22}
...	...	...	...	...	...
2	n_2	X_{2n_2}	Z_2	0	Y_{2n_2}
...	...	...	...	...	...
M	1	X_{M1}	Z_M	0	Y_{M1}
...	...	...	...	...	...
M	n_M	X_{Mn_M}	Z_M	0	Y_{Mn_M}

In what follows, we use $X_i = \{[X_{ij}]_{j=1,\ldots,n_i}, Z_i\}$ to denote the collection of covariates from cluster i (including cluster-level covariates Z_i), $Y_i = [Y_{ij}]_{j=1,\ldots,n_i}$ to denote the vector of outcomes in cluster i, i=1,...,M.

Missing outcome data in CRTs

Cluster	Unit	Covariates		Treatment	Outcome
		Individual-level	Cluster-level	Indicator (A)	
1	1	X_{11}	Z_1	1	Y_{11}
1	2	X_{12}	Z_1	1	Y_{12}
...	...	...	...	...	...
1	n_1	X_{1n_1}	Z_1	1	Y_{1n_1}
2	1	X_{21}	Z_2	0	Y_{21}
2	2	X_{22}	Z_2	0	Y_{22}
...	...	...	...	...	...
2	n_2	X_{2n_2}	Z_2	0	Y_{2n_2}
...	...	...	...	...	...
M	1	X_{M1}	Z_M	0	Y_{M1}
...	...	...	...	...	...
M	n_M	X_{Mn_M}	Z_M	0	Y_{Mn_M}

Multilevel missingness in CRTs

Cluster	Unit	Outcome
1	1	Y_{11}
1	2	?
...	...	...
1	n_1	Y_{1n_1}
2	1	Y_{21}
2	2	Y_{22}
...	...	...
2	n_2	?
...	...	...
M	1	?
...	...	...
M	n_M	Y_{Mn_M}

Cluster	Unit	Outcome
1	1	Y_{11}
1	2	?
...	...	...
1	n_1	Y_{1n_1}
2	1	?
2	2	?
...	...	?
2	n_2	?
...	...	...
M	1	?
...	...	...
M	n_M	Y_{Mn_M}

Missing outcome data in CRTs is common

In a review by Fiero et al. (2016), among 86 CRTs,

- 80 (93%) reported having some missing data at the individual level

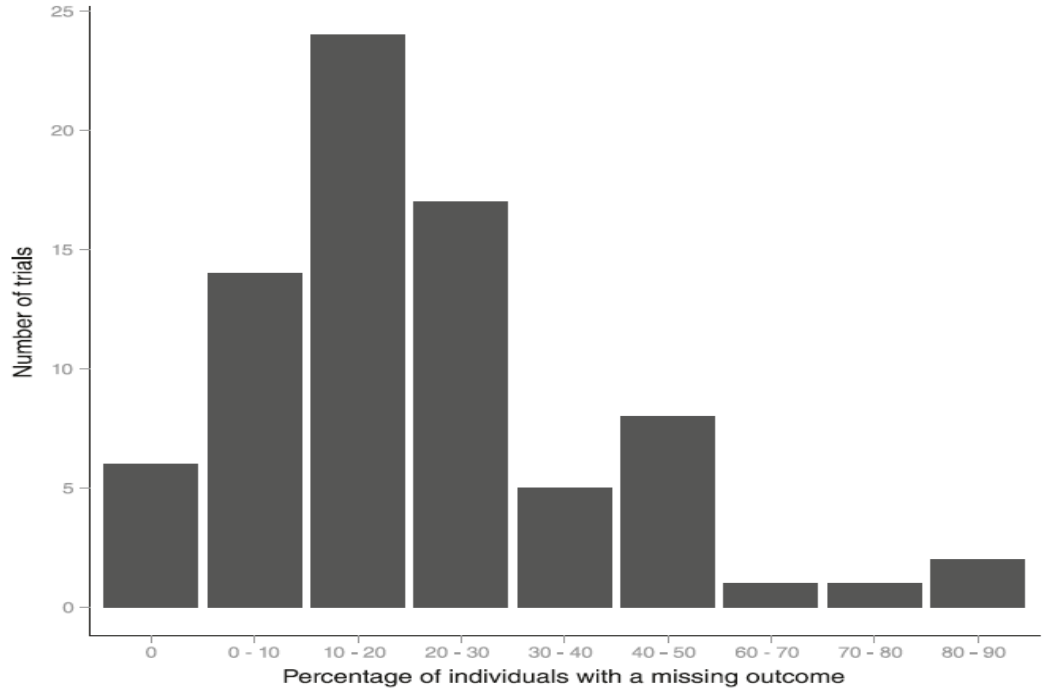


Fig. 2 Distribution of the percentage of individuals with a missing outcome for the 86 trials included in the review

- 27 (31%) reported having whole clusters missing

Data and missingness mechanisms

- Full data: data we would want to collect for all individuals in the sample, $\{(Y_{ij}, X_{ij}, Z_i, A_i), j = 1, \ldots, n_i, i = 1, \ldots, M\}$
- Observed data: data are actually observed, some are missing, $\{(R_{ij}, Y_{ij}R_{ij}, X_{ij}, Z_i, A_i), j = 1, \ldots, n_i, i = 1, \ldots, M\}$.
  - Missingness mechanisms (Rubin 1976):
    - Missing completely at random (MCAR): the probability of missingness does not depend on observed or unobserved information
    - Missing at random (MAR): conditional on the observed data, the probability of missingness is independent of unobserved data
    - Missing not at random (MNAR): neither MCAR nor MAR
- Complete data: data from subsets of individuals without missing data, $\{(R_{ij} = 1, Y_{ij}, X_{ij}, Z_i, A_i), j = 1, \ldots, n_i, i = 1, \ldots, M\}$

Outcome missing mechanisms in CRTs

- MCAR: the missing process is independent of X_i, A_i, and Y_i
- MAR: the missing process can depend on the observed data
- Any component of Y_i can be missing and there is no natural ordering of individual outcomes within a cluster, the missingness pattern can not be assumed as monotone missingness as in the longitudinal data setting.
- A stronger version of MAR is typically assumed.
- Restricted MAR (rMAR): the probability that the outcome for one individual is missing is independent of all outcomes (including the observed outcomes) in the same cluster, conditional on covariates X_i and treatment A_i.

Average treatment effect and ICC

- Interests focus on making an inference about some aspect of the distribution of the full data based on the observed data
- One main goal of CRTs is to estimate the average treatment effect, defined as
$$\Delta = f(E[Y_{ij} \mid A_i = 1], E[Y_{ij} \mid A_i = 0]),$$
where f is a function defining the scale of interest:
  - $f(x, y) = x - y$: difference in means, risk difference
  - $f(x, y) = x(1 - y)/\{(1 - x)y\}$: odds ratio
- Covariates X_{ij} and Z_i are auxiliary variables
- Also of interest is to estimate the intraclass correlation coefficient (ICC): measures the extend to which outcomes are correlated within the same cluster
- Two analytic challenges:
  - Outcome can be missing
  - Data for individuals within each cluster are likely to be correlated.

Commonly-used analysis strategies for CRTs

- Two modeling approaches:
  - Mixed effects models via maximum likelihood estimation (Laird and Ware 1982)
  - Population average models fitted with generalized estimating equation (GEE) approaches (Liang and Zeger 1986)
- Likelihood-based inference in general requires the correct specification of the full likelihood, including the within-cluster correlation structure, which may be hard to specify
- The GEE estimator
  - focuses on population average effects rather than cluster specific effects
  - requires fewer parametric assumptions
  - is robust to misspecification of the correlation structure
- Hubbard et al. (2010) compared population average and mixed models
- Murray et al. (2004) and Turner et al. (2017) reviewed various methodological developments for the analysis of CRTs

The GEE estimator for the average treatment effect

- The standard GEE estimator solves the following estimating equations:

$$0 = \sum_{i=1}^{M} D_i^T V_i^{-1} (Y_i - \boldsymbol{\mu}_i),$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^T$, and $\mu_{ij} = E(Y_{ij} \mid A_i) = g^{-1}(\beta_0 + \beta_A A_i)$, $D_i = \partial \mu_i / \partial \theta^T$ and V_i is a working covariance matrix for Y_i.

- $g(\cdot)$ is a link function. If g is the identity link, β_A represents difference in means for a continuous outcome:

$$\beta_A = E(Y_{ij} \mid A_i = 1) - E(Y_{ij} \mid A_i = 0).$$

- V_i does not need to be correctly specified. Variance for $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_A)^T$ is typically estimated using a sandwich variance estimator (can be obtained by standard software in R 'geepack' or SAS proc GEE).

Analysis strategies in the presence of missing data

- When data are MCAR, the standard GEE estimator based on complete data is consistent and asymptotically normal
- When data are MAR, the standard GEE estimator based on complete data may yield biased estimates
- Potential solutions under rMAR:
  - Multiple imputation (MI-GEE): requires specification of an imputation model for $E(Y_{ij} \mid X_i, A_i)$
  - Inverse probability weighting (IPW-GEE): requires specification of a propensity score model for $P(R_{ij} = 1 \mid X_i, A_i)$
  - Augmented inverse probability Weighting (AIPW-GEE): requires specification of a propensity score model for $P(R_{ij} = 1 \mid X_i, A_i)$ and an outcome model for $E(Y_{ij} \mid X_i, A_i)$
  - "Multiply robust" AIPW-GEE: allows specification of a set of models for the propensity score model and a set of models for the outcome model, requires one model to be correctly specified

Multi-level multiple imputation (MMI-GEE)

- Steps:
  - Missing values are imputed multiple times using a full-parametric model
  - Resulting complete data sets are analyzed using a standard GEE approach
  - Results are then combined across multiple imputed datasets (Rubin 2004)
- For CRTs, two practical considerations:
  - The imputation model must properly account for the multi-level data structure
  - Use treatment arm specific imputation model
- For more MI for CRTs, please see Dr. Rebecca Andridge's talk at: `https://prevention.nih.gov/education-training/methods-mind-gap/multiple-imputation-methods-group-based-interventions`

Inverse probability weighting (IPW-GEE)

- Reweighting complete cases according to the probability of being observed (Robins et al., 1995) so that an individual with complete data is considered representative of him/herself as well as a number of similar subjects that had dropped out from the study

$$0 = \sum_{i=1}^{M} D_i^T V_i^{-1} W_i [Y_i - \boldsymbol{\mu}_i],$$

where $W_i = diag[R_{ij}/\hat{\pi}_{ij}]_{j=1,\ldots,n_i}$, $\hat{\pi}_{ij}$ can be obtained by fitting a binary response model that regresses the indicator R_{ij} on functions of A_i and X_i, referred to as the propensity score model

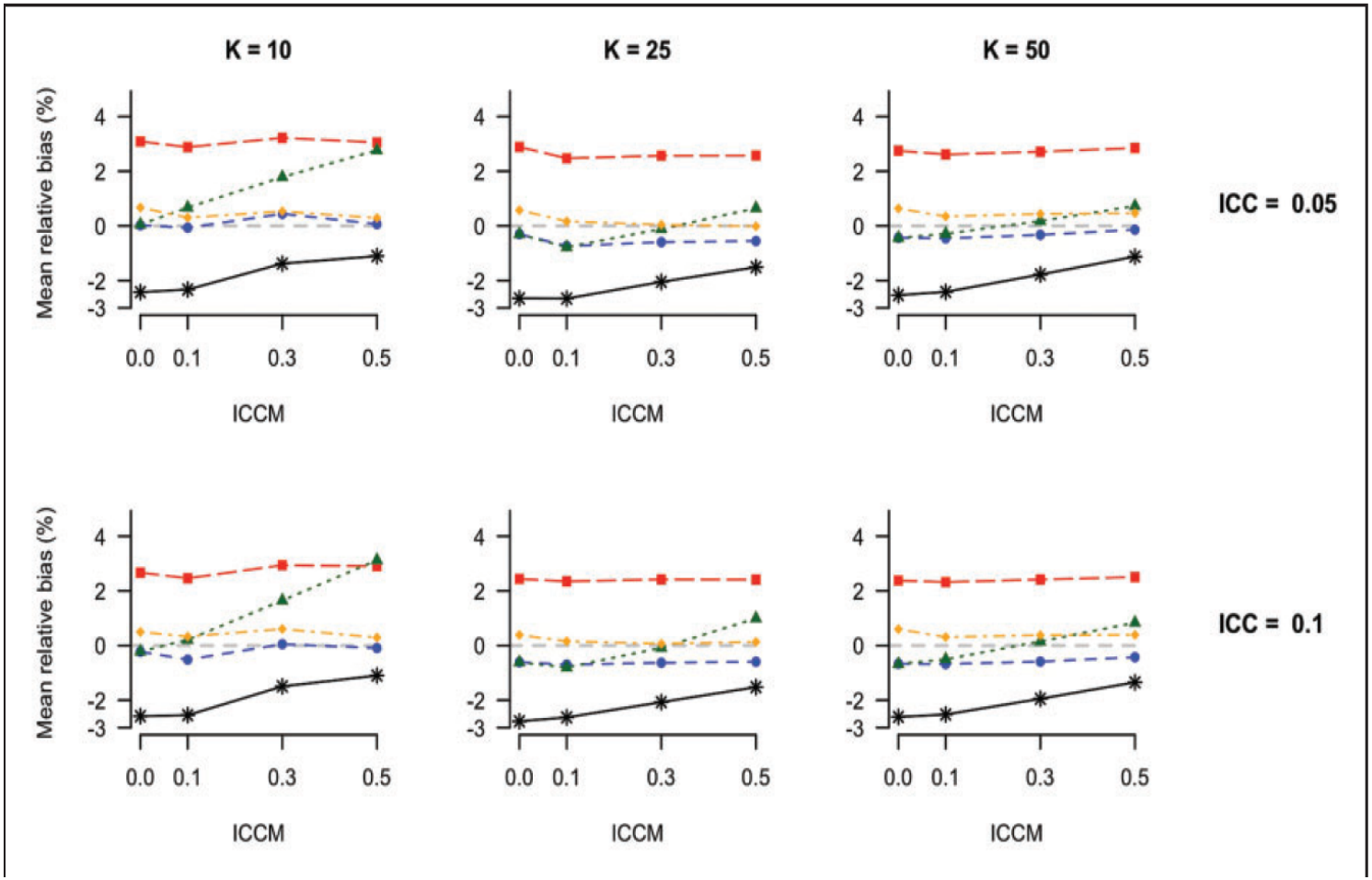
- The resulting estimator is consistent and asymptotically normal provided that the propensity score model is correctly specified:

$$\pi_{ij}(X_i, A_i; \eta_W) = P(R_{ij} = 1 \mid X_i, A_i),$$

for some η_W.

MMI-GEE vs. IPW-GEE

Simulation studies report comparable performance of MMI-GEE and IPW-GEE in CRTs with missing binary outcome data (Turner et al., 2020)



black: complete case analysis; red: adjusted complete case analysis; orange: MMI-GEE; blue: IPW-GEE; teal: IPW-GEE accounting for clustering when estimating the weights

Augmented IPW-GEE

In practice, we may not know whether the propensity score model is correct specified. Can augment the estimating equation with a term that relates the outcome to covariates and treatment (Prague et al., 2016).

$$0 = \sum_{i=1}^{M} [D_i^T V_i^{-1} W_i (Y_i - B_i) + \sum_{a=0,1} p^a (1-p)^{1-a} D_i^T V_i^{-1} (B_i - \boldsymbol{\mu}_i)],$$

where $p = P(A_i = 1)$, $B_i(X_i, A_i = a; \eta_B)$ is referred to as the outcome model, it is correctly specified when

$$B_{ij}(X_i, A_i = a; \eta_B) = E(Y_{ij} \mid X_i, A_i = a)$$

for some parameters η_B .

Enjoys the doubly robust property: the resulting AIPW-GEE estimator is consistent and asymptotically normal if either the propensity score model or the outcome model is correctly specified.

Estimating the intraclass correlation coefficient (ICC)

- Second-order estimating equations based on pairs of observations can be used to estimate the intraclass correlation coefficient (Zhao and Prentice 1990, Yan and Fine 2004, Yi and Cook 2002)

- Chen et al. (2020) adopt a specific parametrization that targets the treatment-specific ICC ρ_i

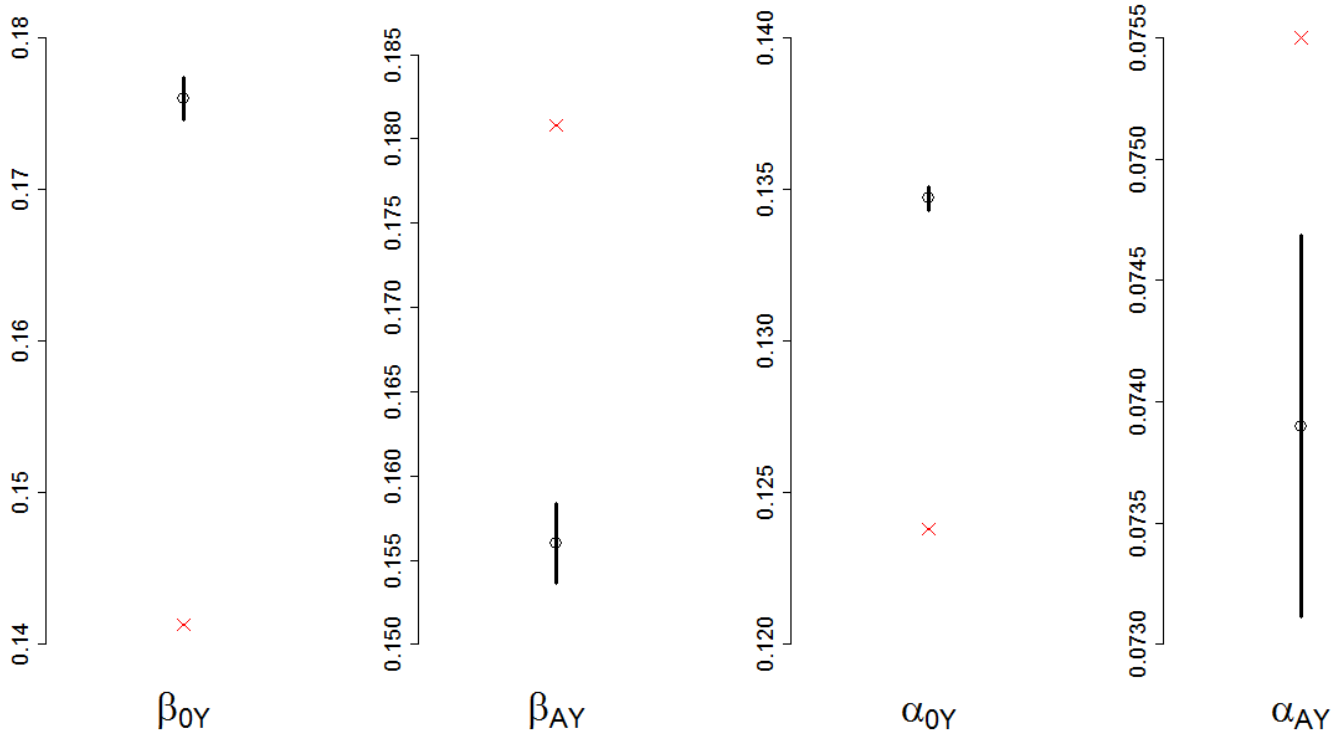
$$0 = \sum_{i=1}^{M} D_i^T V_i^{-1}(Y_i - \boldsymbol{\mu}_i), \quad \text{logit}(\mu_i) = \beta_{0Y} + \beta_{AY} A_i$$

$$0 = \sum_{i=1}^{M} \widetilde{D}_i^{-1} \widetilde{V}_i^T (\mathcal{E}(Y_i) - \boldsymbol{\rho}_i), \quad \text{atanh}(\rho_i) = \alpha_{0Y} + \alpha_{AY} A_i$$

$$\mathcal{E}(Y_i) = \left[\frac{(Y_{ij} - \mu_i)(Y_{ij'} - \mu_i)}{\mu_i(1 - \mu_i)} \right]_{j<j'}$$

Complete case analysis leads to bias

- Simulate correlated binary data for outcome Y_{ij} and missingness indicator R_{ij} with number of clusters $M = 2000$ and cluster sizes $n_i \in \{81, \cdots, 140\}$



× True value —— Complete Case GEE2

IPW-GEE2

- IPW-GEE1:

$$Y_{ij} - \mu_i \mapsto \frac{R_{ij}}{\pi_{ij}}(Y_{ij} - \mu_i)$$

  - π_{ij} is the propensity score model for $P[R_{ij} = 1 | X_i, A_i]$

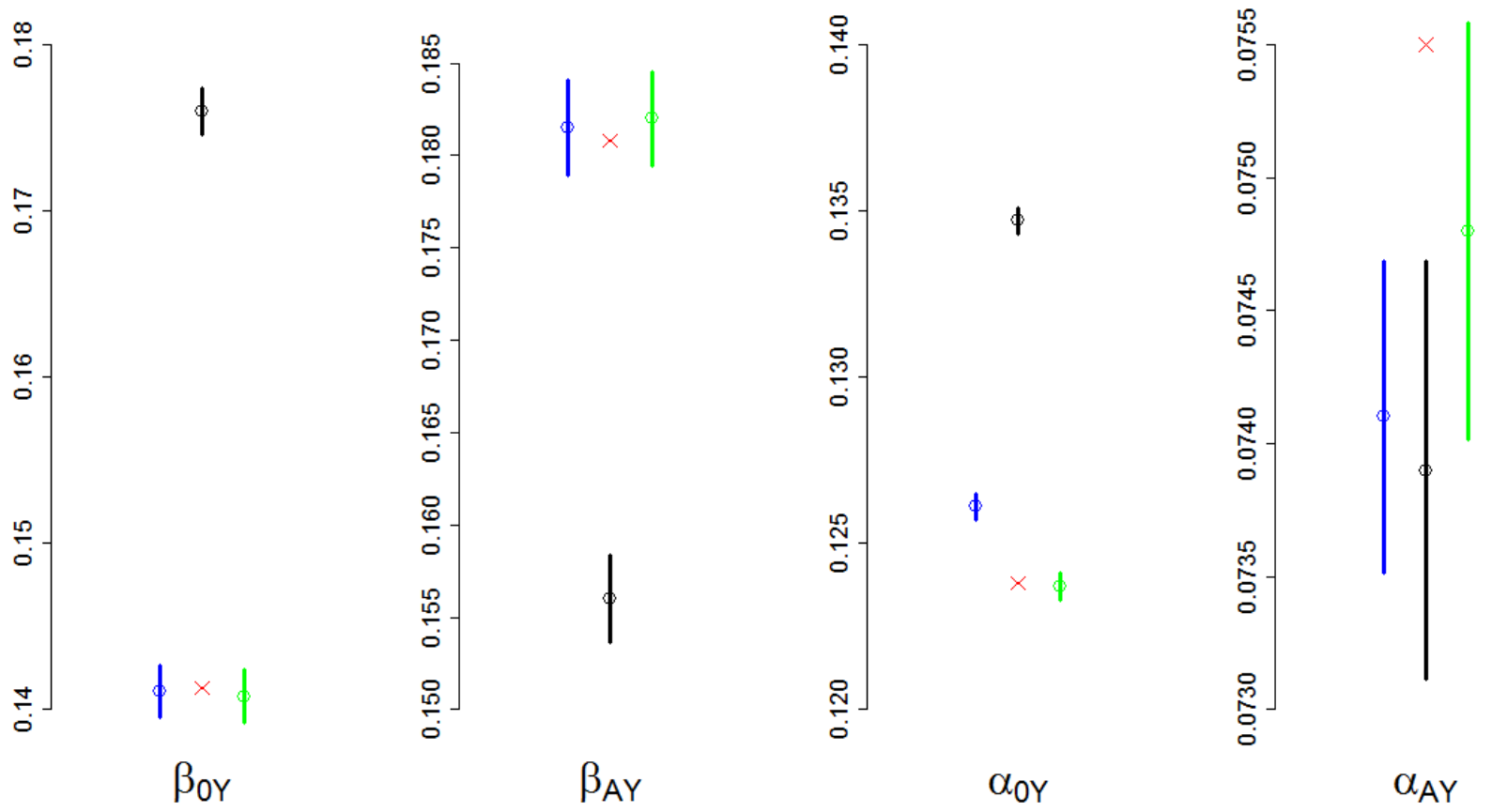
- IPW-GEE2:

$$\frac{(Y_{ij} - \mu_i)(Y_{ij'} - \mu_i)}{\mu_i(1 - \mu_i)} - \rho_i \mapsto \frac{R_{ij}R_{ij'}}{\eta_{ijj'}} \left[\frac{(Y_{ij} - \mu_i)(Y_{ij'} - \mu_i)}{\mu_i(1 - \mu_i)} - \rho_i \right]$$

  - $\eta_{ijj'}$ is a model for $E[R_{ij}R_{ij'} | X_i, A_i]$, referred to as the *second-order propensity scores*

- Ignoring "correlated missingness" $[\eta_{ijj'} = \pi_{ij}\pi_{ij'}]$ can lead to biased $\widehat{\boldsymbol{\alpha}}_Y$.

IPW-GEE2 – PS models correctly specified



× True value —— Complete Case GEE2 $\mathcal{G}_1(\mathbf{R})$ IPW-GEE2 —— $\mathcal{G}_2(\mathbf{R})$ IPW-GEE2

Doubly robust estimator

- Can similarly derive the DR estimator, which is consistent and asymptotically normal under correct specification of either the outcome model or the propensity score model

- Denote $\boldsymbol{\kappa} = (\beta_Y, \alpha_Y, \eta_W, \eta_B)$ and

$$\Psi_i(\boldsymbol{\kappa}) = \begin{pmatrix} \widetilde{\Phi}_i^Y(A_i, X_i, R_i, \beta_Y, \alpha_Y, \eta_W, \eta_B) \\ \mathbf{S}_i^W(A_i, X_i, \eta_W) \\ \mathbf{S}_i^B(A_i, X_i, \eta_B) \end{pmatrix}$$

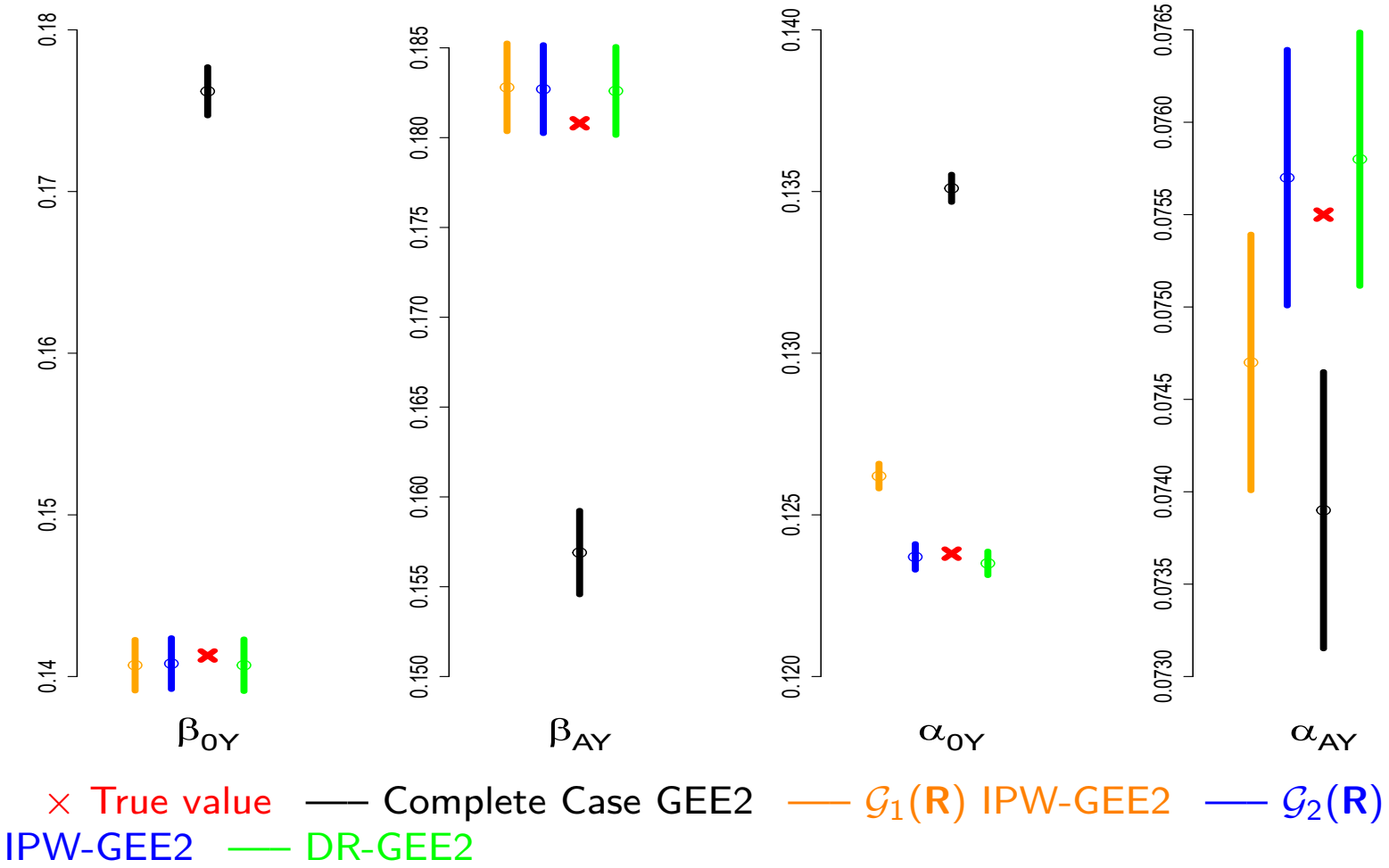
- By standard results for M-estimators, $\sqrt{M}(\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}_0) \xrightarrow{\mathcal{D}} N(0, \Gamma^{-1}\Delta(\Gamma^{-1})^\intercal)$, where

$$\Delta = E[\Psi(\boldsymbol{\kappa}_0)\Psi(\boldsymbol{\kappa}_0)^\intercal] \quad \text{and} \quad \Gamma = E\left[\frac{\partial \Psi(\boldsymbol{\kappa}_0)}{\partial \boldsymbol{\kappa}^\intercal}\right]$$

from which we can extract components corresponding to $(\widehat{\beta}_Y, \widehat{\alpha}_Y)$.

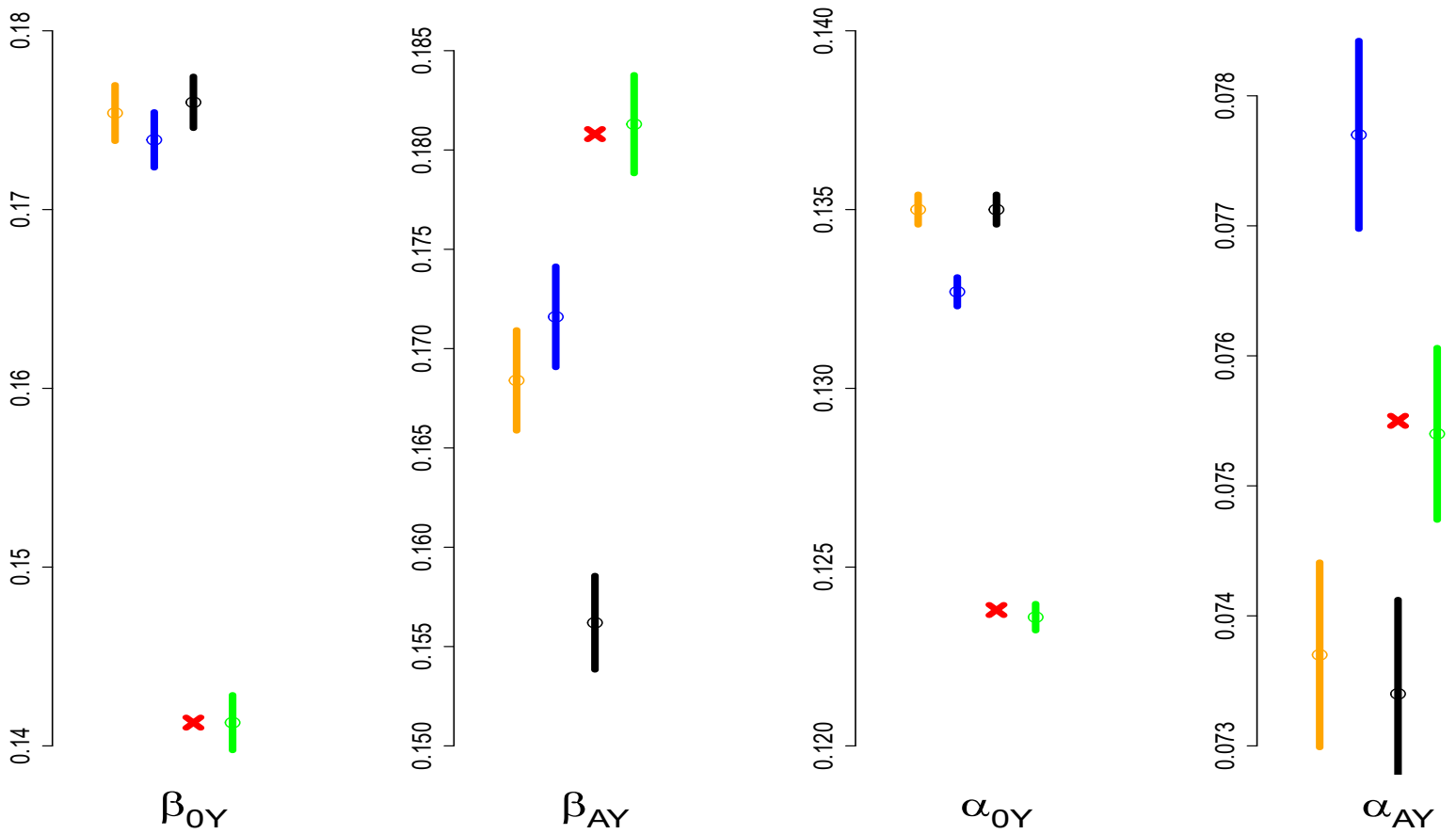
Simulation results: Both propensity score and outcome models correct

$M = 2000$ with $n_i \sim \text{Unif}\{80, \cdots, 140\}$



Simulation results: Only the outcome model correct

$M = 2000$ with $n_i \sim \text{Unif}\{80, \cdots, 140\}$



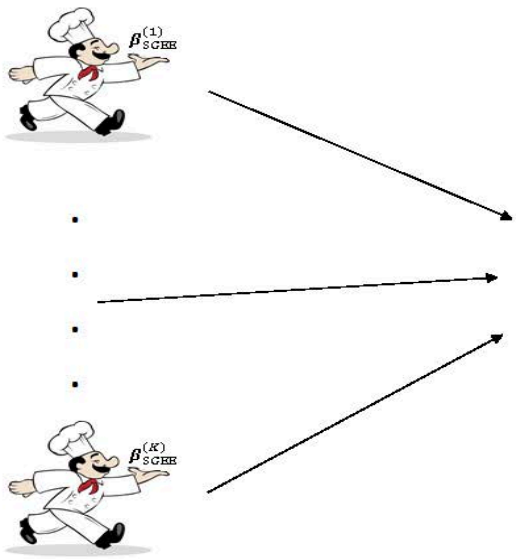
× True value —— Complete Case GEE2 —— $\mathcal{G}_1(\mathbf{R})$ IPW-GEE2 —— $\mathcal{G}_2(\mathbf{R})$ IPW-GEE2 —— DR-GEE2

Computational challenges with fitting GEEs

- Computational challenges in solving GEEs has been noted by many (Carey et al., 1993; Yan and Fine 2004)
- Second-order GEEs include an extra set of estimating equations, involving all possible pairs of observations
- The computing complexity increases quadratically as the cluster sizes increase
- Solving GEEs with large cluster sizes becomes difficult due to both convergence and memory allocation issues
- Chen et al. (2020) proposes stochastic algorithms to alleviate this issue: at each Newton-Raphson iteration, only use a subsample

Parallel stochastic GEE algorithm

- Stochastic GEE allow faster computation in each iteration, but each iteration is not as informative due to the induced missingness
- Instead of one chef cooking ten meals, hire ten chefs to cook each meal



$$\beta_{\mathrm{ParSGEE}} = \frac{1}{K}\sum_{k=1}^{K} \beta_{\mathrm{SGEE}}^{(k)}$$

- Improve on convergence by intrinsically incorporating information in its multistart search

AIPW-GEE: A multiply robust version

- AIPW-GEE estimator is consistent and asymptotically normal if either the propensity score model for the outcome missingness or the covariate-conditional mean outcome model is correctly specified

- A multiply robust version (Rabideau et al., 2024): Consider specification of multiple propensity score models and multiple covariate-conditional mean outcome models, the resulting estimator is consistent and asymptotically normal as long as one model is correctly specified, for example:

Set	Label	Model	Correct
PS	0	$\text{logit}\{\pi(X,A)\} = \theta_I + \theta_A A + (X_1, X_2, X_2^2, X_2^3, e^{X_3})\boldsymbol{\nu}$	Yes
Model	1	$\text{logit}\{\pi(X,A)\} = \theta_I^{(1)} + \theta_A^{(1)} A + (1_{X_1>0}, 1_{X_2>-1}, X_3)\boldsymbol{\nu}^{(1)}$	No
	2	$\text{logit}\{\pi(X,A)\} = \theta_I^{(2)} + \theta_A^{(2)} A + X_2 \nu^{(2)}$	No
Outcome	0	$E(Y\|X,A) = \beta_I + \beta_A A + (X_1, X_2, X_2^2, X_2^3, e^{X_3})\boldsymbol{\zeta}$	Yes
Model	1	$E(Y\|X,A) = \beta_I^{(1)} + \beta_A^{(1)} A + (1_{X_1>0}, 1_{X_2>-1}, X_3)\boldsymbol{\zeta}^{(1)}$	No
	2	$E(Y\|X,A) = \beta_I^{(2)} + \beta_A^{(2)} A + X_2 \zeta^{(2)}$	No

- An R package for fitting a MR-GEE is available at `https://github.com/djrabideau/mrgee`

Multiple-level IPW-GEE

- So far we focus on missingness at the individual-level, i.e., no empty clusters
- Entire clusters (or subclusters) can be missing in CRTs with a multi-level structure
- An example of a CRT with multiple level missingness:
  - A CRT was conducted to evaluate if proactive community care management (pro-CCM) is effective in reducing malaria burden in rural endemic area of Madagascar, twenty-two fokontanies (smallest administrative units) were randomized to pro-CCM or conventional integrated community case management (Ratovoson et al. 2022)
  - The study participants were nested in households, which were nested in each fokontany
  - About 24% of study participants and 22% of the households were lost to follow-up due to moving away, absence, death, or refusal to participate

Multi-level missingness process

- In addition to the individual-level missingness indicators R_{ij}, introducing a cluster-level missingness indicator C_i
- A cluster-level missingness model: $\lambda(A_i, Z_i; \gamma) = P(C_i = 1 \mid A_i, Z_i)$
- An individual-level missingness model: $\pi(A_i, Z_i, X_{ij} \mid C_i = 1; \eta) = P(R_{ij} = 1 \mid C_i = 1, A_i, Z_i, X_{ij})$
- Weights given by

$$W_i = diag[\frac{R_{ij} C_i}{\pi_{ij} \lambda_i}]_{j=1,\ldots,n_i}.$$

- Can specify a set of models for λ and a set of models for π

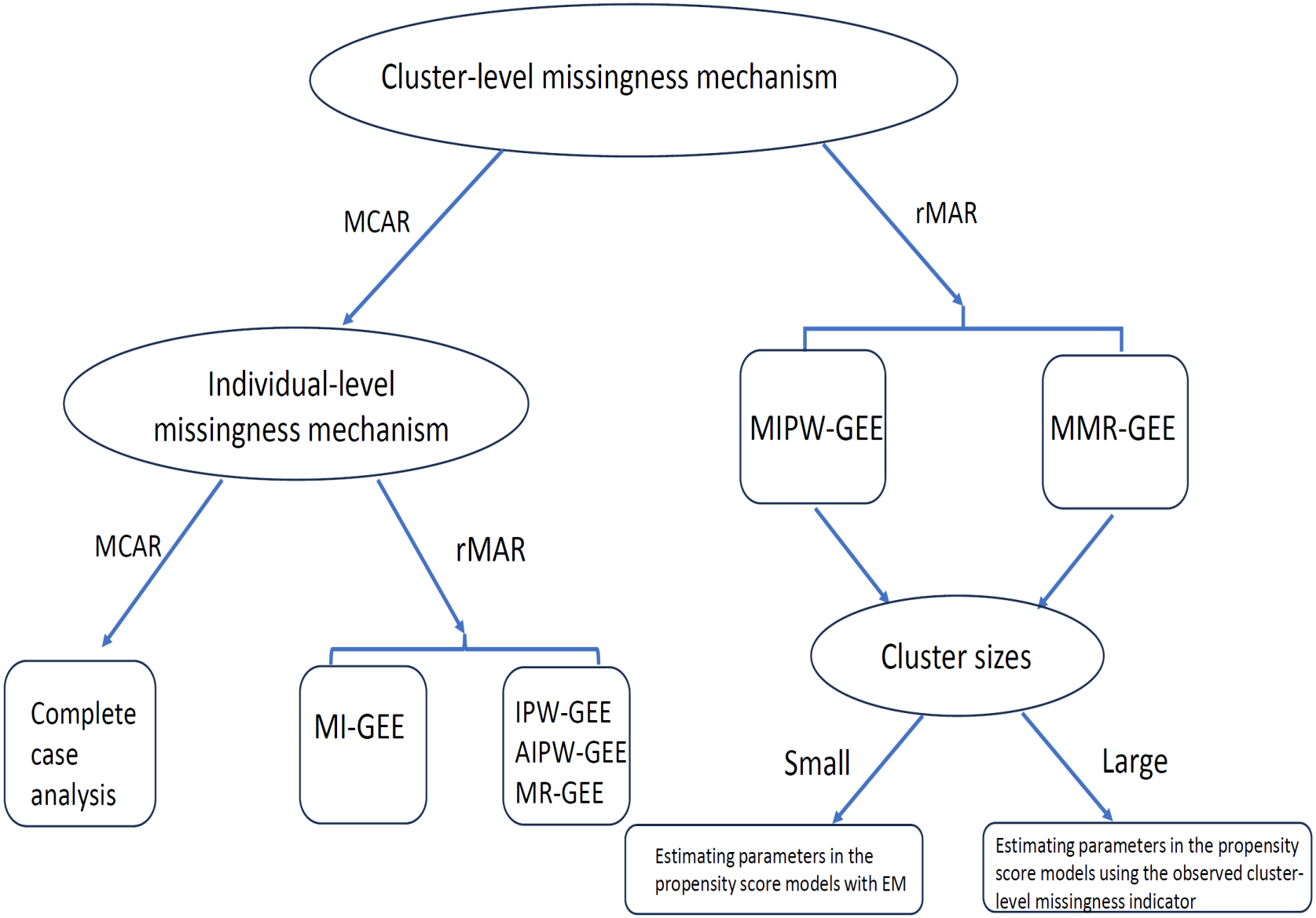
Misclassification of cluster-level missingness indicator

When cluster sizes are small, the observed cluster-level missingness indicator may be misclassified in the sense that, when no outcomes are observed in a cluster (e.g., cluster 2), we may not know whether it is due to the cluster being withdrawn or due to all individual outcomes being missing

Cluster	R_{ij}	C_i^O	C_i
1	0	0	0
	0	0	0
	0	0	0
2	0	0	1
	0	0	1
	0	0	1
3	1	1	1
	0	1	1
	0	1	1

Incorporate an Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to learn about the true C_i in estimation of nuisance parameters in the individual-level and cluster-level propensity score models

Summary of modeling and estimation methods



References

- Chen T, Tchetgen Tchetgen EJ, Wang R (2020). A stochastic second-order generalized estimating equations approach for estimating association parameters. Journal of Computational and Graphical Statistics, 29: 3, 547-561.
- Carey V, Zeger SL, and Diggle P (1993). Modelling multivariate binary data with alternating logistic regressions. Biometrika, 80, 517-526.
- Dempster AP, Laird NM and Rubin DB (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B 39, 1-22.
- Fiero MH, Huang S, Oren E, Bell ML (2016). Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. Trials. 17: 72.
- Hubbard, AE, Ahern, J, Fleischer, NL, Lann, MV, Lippman, SA, Jewell N, Brucker T, Satariano, WA (2010). To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology 21(4): 467-474.
- Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics, 1982; 38: 963-974.
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika, 1986; 73: 13-22.
- Murray DM, Varnell SP, Blitstein JL (2004). Design and analysis of group-randomzied trials: a review of recent methodological developments. 94(3): 423-32.
- Paik, MC (1997). The generalized estiamting equation approach when data are not missing completely at random. Journal of the American Statistical Association; 92, 1320-1329.

References

- Prague, M, Wang, R, Stephens, A, Tchetgen Tchetgen, E, DeGruttola, V (2016). Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes. Biometrics, 72(4), 1066-1077.
- Rabideau DJ, Li F and Wang R (2024). Multiply robust generalized estimating equations for cluster randomized trials with missing outcomes. Statistics in Medicine, provisionally accepted.
- Ratovoson et al. (2022). Proactive community care management decreased malaria prevalence in rural Madagascar: results from a cluster randomized trial. BMC Med. 20(1): 322.
- Robins JM, Rotnitzky A and Zhao LP (1995). Analysis of semiparametric rergession-models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association; 90: 106-121.
- Rubin DB. Inference and missing data. Biometrika 63, 581-592.
- Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken: Wiley-Interscience, 2004.
- Turner, EL, Prague M, Gallis JA, Li F, Murray DM (2017). Review of recent methodological developments in group-randomized trials: Part 2 Analysis. 107(7): 1078-1086.
- Turner, EL, Yao L, Li F, and Prague M (2020). Properties and pitfalls of weighting as an alternative to multilevel multiple imputation in cluster randomized trials with missing binary outcomes under covariate-dependent missingness. Stat Methods Med Res. 29(5): 1338-1353.

References

- Yan J and Fine J (2004). Estimating equations for association structures. Statistics in Medicine, 23(6), 859-874.
- Yi GY and Cook RJ (2002). Marginal methods for incomplete longitudinal data arising in clusters. Journal of the American Statistical Association, 97:460, 1071-1080.
- Zhao LP and Prentice RL (1990). Correlated binary regression using a quadratic exponential model. Biometrika, 77, 642-548.