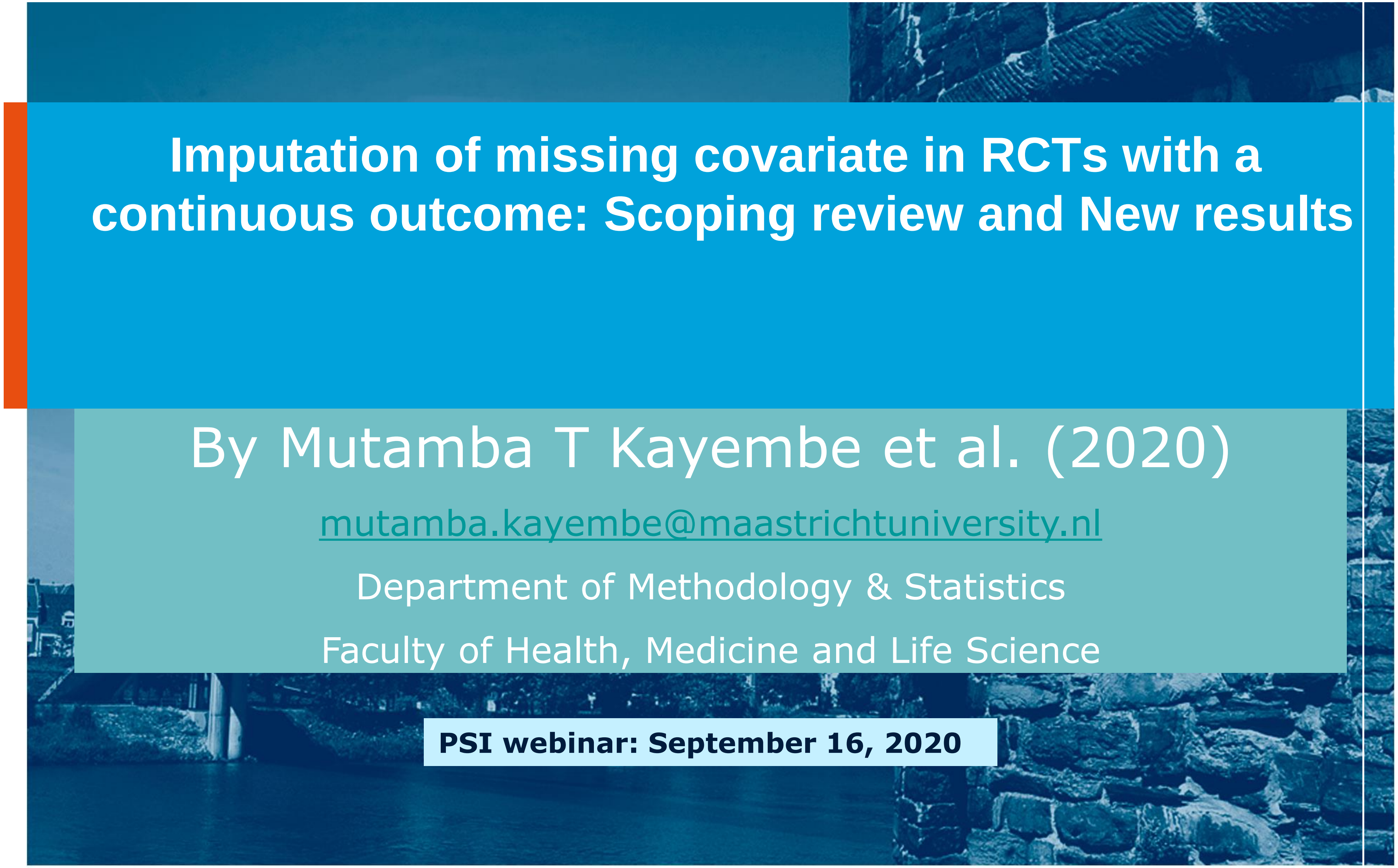
Maastricht University *Leading in Learning!*

CAPHRI School for Public Health and Primary Care



Imputation of missing covariate in RCTs with a continuous outcome: Scoping review and New results

By Mutamba T Kayembe et al. (2020)

mutamba.kayembe@maastrichtuniversity.nl

Department of Methodology & Statistics

Faculty of Health, Medicine and Life Science

PSI webinar: September 16, 2020

Outline

1. Problem statement
2. Method: *Overview*
3. Method: *Simulation setup*
4. Results
5. Discussion

Problem Statement

- Optimal (suboptimal) methods for handling missing covariates in nonrandomized studies should not be expected to necessarily be optimal (suboptimal) in randomized studies.
- **Example: The belief that multiple imputation (MI)** is the method of choice for handling missing covariates is generally based on nonrandomized studies.
- **What about in randomized controlled trials (RCTs): Is MI still the method of choice for handling missing covariates in RCTs?**

Method: *overview*

- Scope review the literature on handling missing covariates in RCTs with a continuous outcome to identify the gaps that need to be filled;
- Gap focused on: Imputation of missing binary covariate,
  - Comparing MI vs. simple alternative methods in RCTs;
- Do so through simulation under a wide range of scenarios;
- Distinguish situations with pre- and post-randomization covariate (but measured before treatment);
  - **Hence**: *more missingness mechanisms than in previous studies*;
  - **Note**: *post-randomization covariate is not affected by treatment, only its missingness*.

Simulation setup: *Analysis of interest*

- *Primary focus*: Linear regression model with two covariates (*T* and *Z*):

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Z_i + \varepsilon_i; \quad i = 1, \dots n$$

  - ❑ T is the treatment indicator, β_1 the treatment effect of interest, and Z the pretest of the outcome Y;
  - ❑ Missingness can occur in Z;

- *Extension (E)*: Cox PH regression model with two covariates (*T* and *Z*):

$$h_X(x|T,Z) = h_0(x) ex\,p(\beta_1 T + \beta_2 Z)\,; \quad where$$

  - ❑ X(survival times) based on Weibull distribution: $h_X(x) = \lambda_X k x^{k-1} \exp(\beta_1 T + \beta_2 Z)$, with
    - λ_X and k as scale and shape parameters, respectively.
  - ❑ Random censoring times based on Weibull distribution: $h_C(x) = \lambda_C k x^{k-1}$

Simulation setup: *Generating complete data*

- Parallel group trial data of sample size n allocated, randomly and evenly, to two treatment groups, (*T=0*) and (*T=1*), as follows:
  - **Sample size**: Small (*n=100*) and large (*n=400*);
  - **Covariate**: *Z*~ Bernoulli, with *P* (*Z=0*) = *P* (*Z=1*);
  - **Treatment assignment**:
    $P(T=0|_{Z=0}) = P(T=0|_{Z=1}) = P(T=1|_{Z=0}) = P(T=1|_{Z=1})$
  - **Outcome**:

    $Y_i = \beta_0 + \beta_1 T_i + \beta_2 Z_i + \varepsilon_i$; where:

    $\varepsilon_i \sim N(0,1)$;

    $\beta_0 = 0$; and $(\beta_1, \beta_2) = (1,1); (1,2); (2,1); (2,2)$
- And...

Simulation setup: *Creating missingness*

- **Create missingness on Z using the model**:

  $logit\{Pr(R=1)\} = \alpha_0 + \alpha_1 Z + \alpha_2 T + \alpha_3 Y + \alpha_4 ZT$; where:

  $R = 0$ if Z is missing and $R = 1$ if Z is observed.

- **Five missingness mechanisms considered**:
  - ❑ **Case 1**: **Z** measured pre-or post-randomization (but before treatment (*T*))**:**
    - **MCAR**: Missing completely at random
    - **MNAR1**: Missingness of *Z* depends on *Z*
  - ❑ **Case 2**: **Z** measured post-randomization (but before treatment (*T*)):
    - **MAR**: Missingness of *Z* depends on *T*
    - **MNAR2**: Missingness of *Z* depends on additive effect of *Z* and *T*
    - **MNAR3**: Missing of *Z* depends on additive effect of *Z*, *T* and *ZT*

Simulation setup: *Overview of all the simulation conditions*

Table 1: The simulation conditions (2×4×3×8 = 192) obtained by combining the parameters.

Sample size n: 100; 400. **1500 datasets for each scenario**				
Treatment and covariate effects (β_1, β_2): **(1, 1); (1, 2); (2, 1); (2, 2)**				
Missingness rates: **20%; 40%; 60%**				
Missingness mechanisms:				
MCAR	MAR	MNAR1	MNAR2	MNAR3
$\alpha_0 \neq 0$	$\alpha_0 \neq 0$, and $\alpha_2 = 0.5$	$\alpha_0 \neq 0$, and $\alpha_1 = \begin{cases} 0.5 \\ 2 \end{cases}$	$\alpha_0 \neq 0$, and $(\alpha_1, \alpha_2) =$ $\begin{cases} (0.5,\ 0.5) \\ (0.5,\ 2) \end{cases}$	$\alpha_0 \neq 0$, and $(\alpha_1, \alpha_2, \alpha_4) =$ $\begin{cases} (0.5,\ 0.5,\ 1) \\ (2,\ 0.5,\ 1) \end{cases}$

Note: For each missingness mechanism, the α’s not shown were set to 0

Simulation setup: *Imputing the missing data and analyzing the imputed data*

1. **Imputation stage**: Impute the missing data, using the method at hand (**Meth**; for instance, **mean imputation**)

2. **Analysis stage**: Apply the analysis of interest on each imputed dataset and produce:

   - **The treatment effect estimate: $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$**;
   - **The standard error (SE) of $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$**;

3. Repeat 1 and 2 several times (=1500) and produce the performance criteria

Simulation setup: *Performance criteria*

1) Bias of $\hat{\beta}_1$
2) Coverage of 95% CI
3) Relative precision (RP) of $\hat{\beta}_1$
4) Relative bias (RB) of estimated SE
5) Relative precision (RP) of estimated SE

❖ **Note**: **4) and 5)** ***are not shown here due to time constraints***

Simulation setup: *Methods compared*

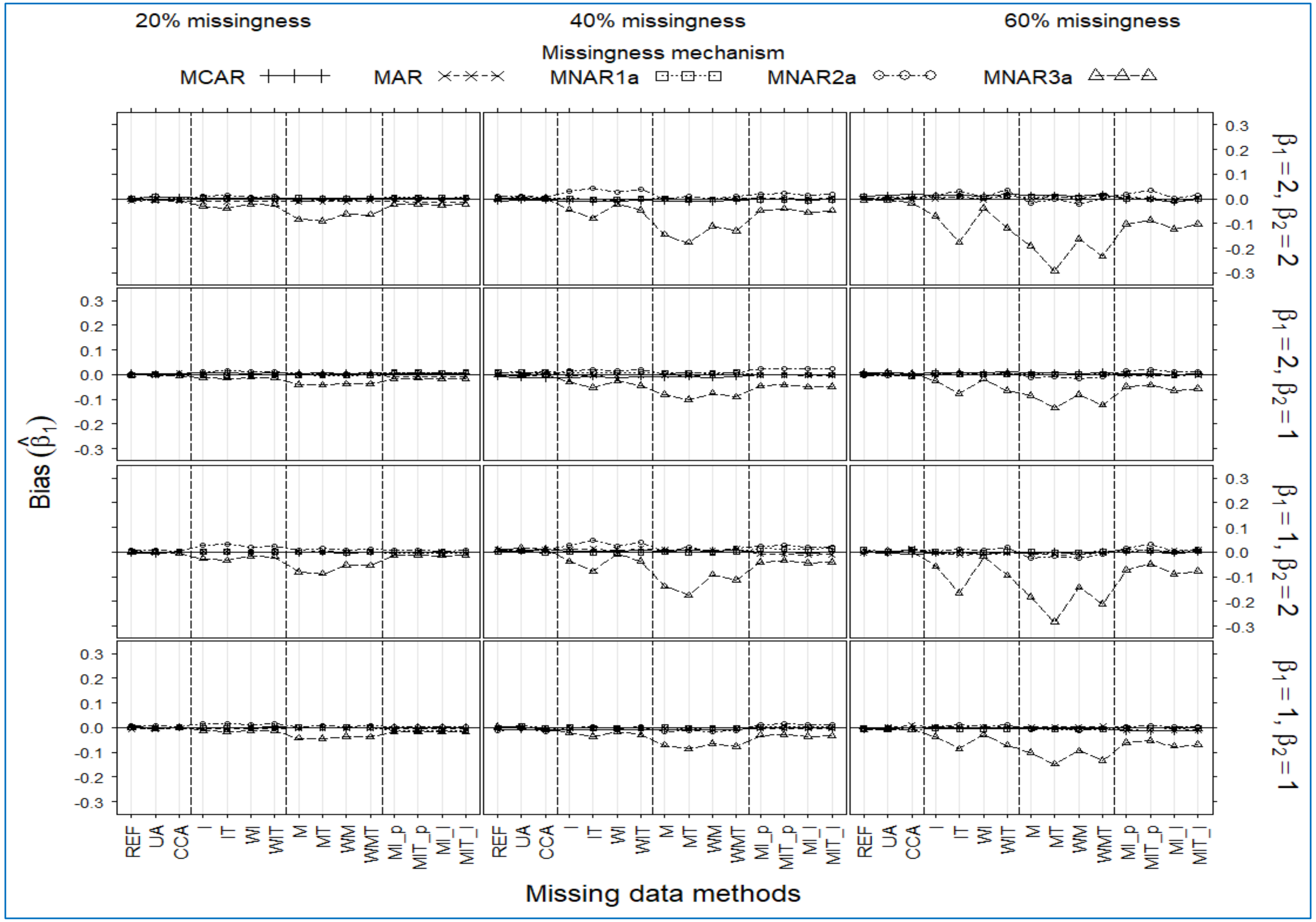
1) **No imputation**:
    - Analysis on complete data: (**REF**)
    - Unadjusted analysis: (**UA**)
    - Complete-case analysis: (**CCA**)

2) **Mean imputation**:
    - Across treatment *T:* (**I**)
    - Per treatment *T* : (**IT**)
    - Weighted, across treatment *T:* (**WI**)
    - Weighted, per treatment *T:* (**WIT**)

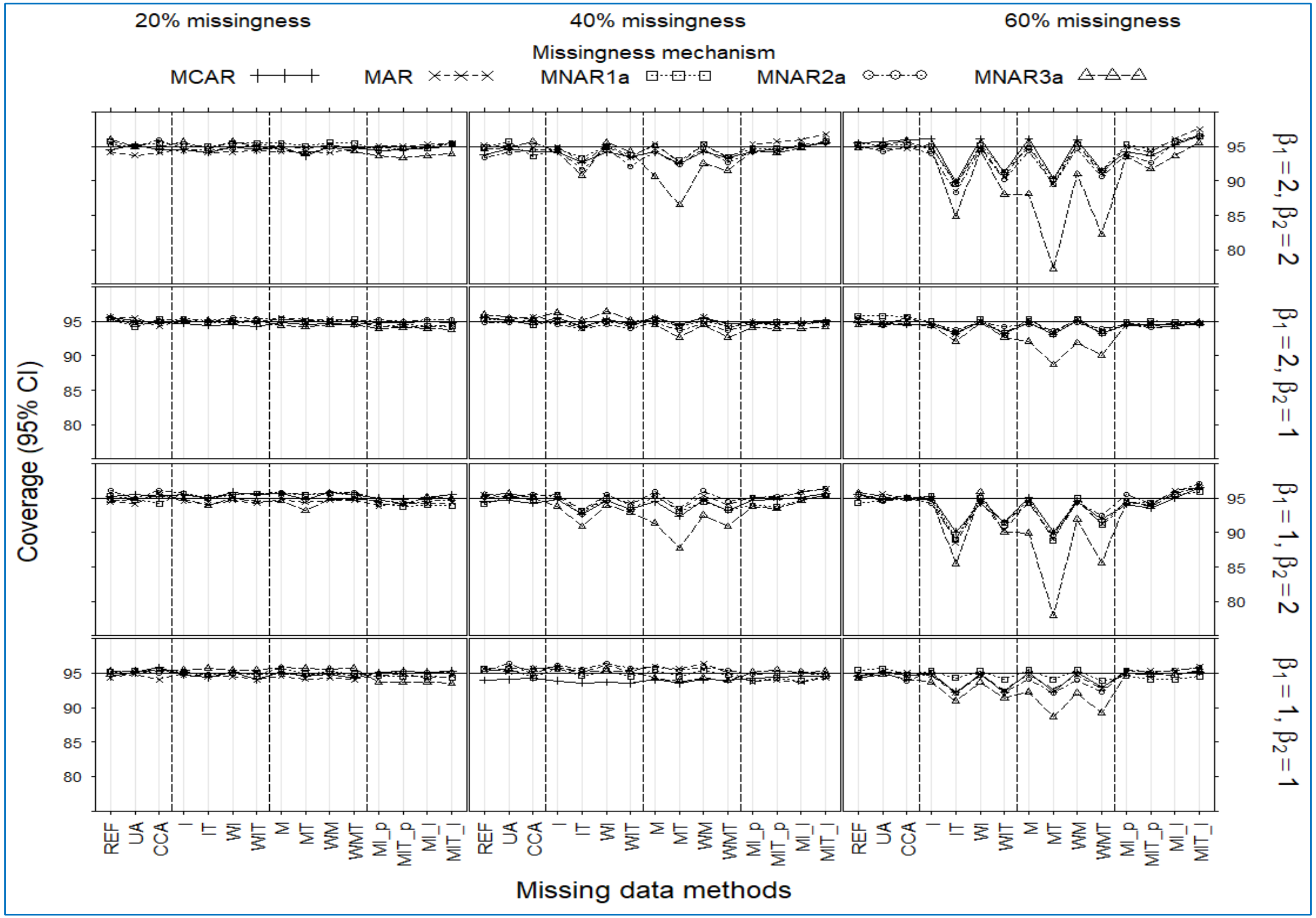
3) **Missing-indicator method**:
    - Across *T*: (**M**)
    - Per *T*: (**MT**)
    - Weighted, across *T*: (**WM**)
    - Weighted, per *T*:(**WMT**)

4) **Multiple imputation (MI)**:
    - Across *T* with predictive mean matching (PMM): (**MI_p**)
    - Per *T* with PMM: (**MIT_p**)
    - Across *T* with logistic regression: (**MI_l**)
    - Per *T* with logistic regression: (**MIT_l**)

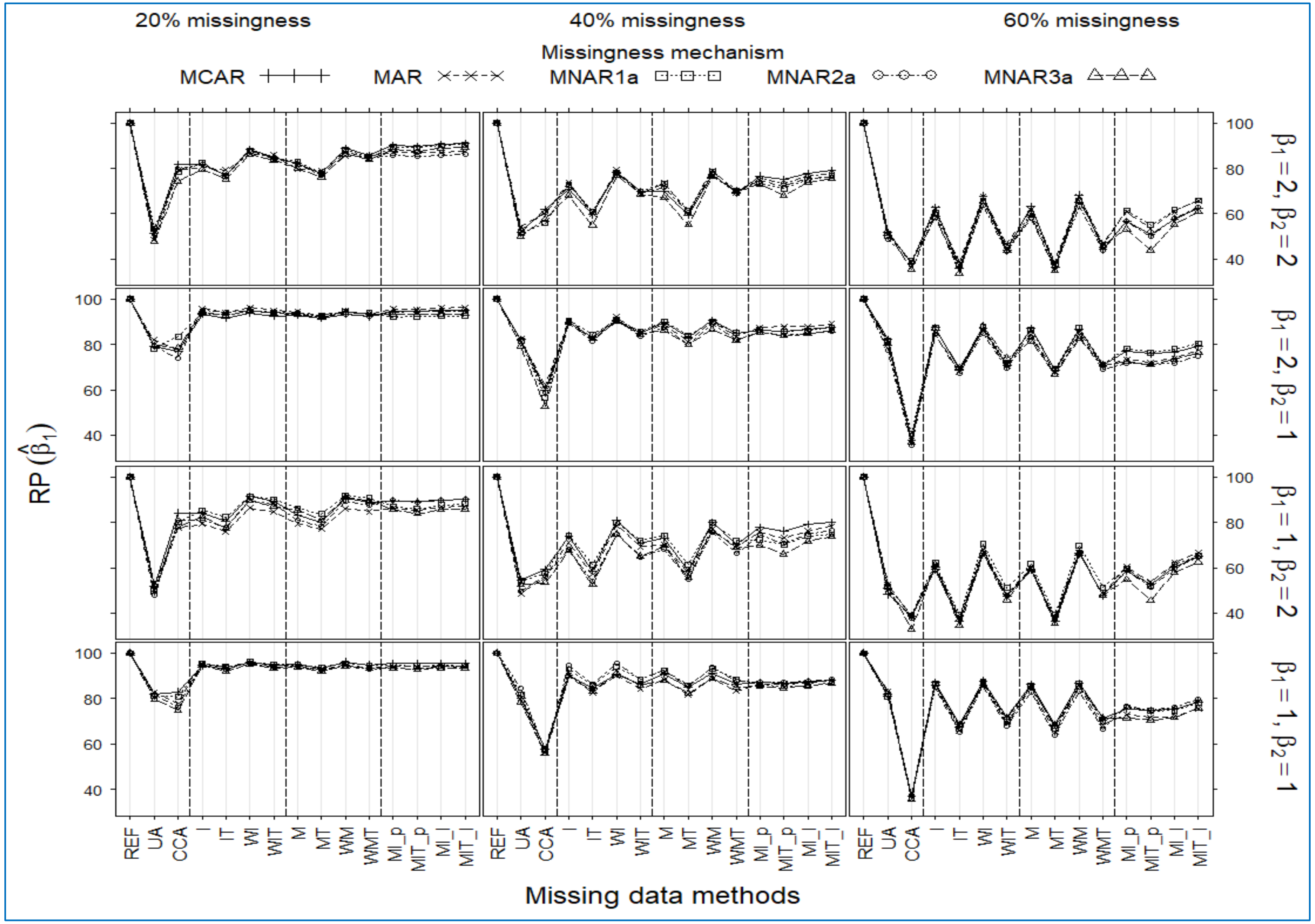
Simulation results for continuous outcome: *Bias of* $\hat{\beta}_1$ (Figure 1)



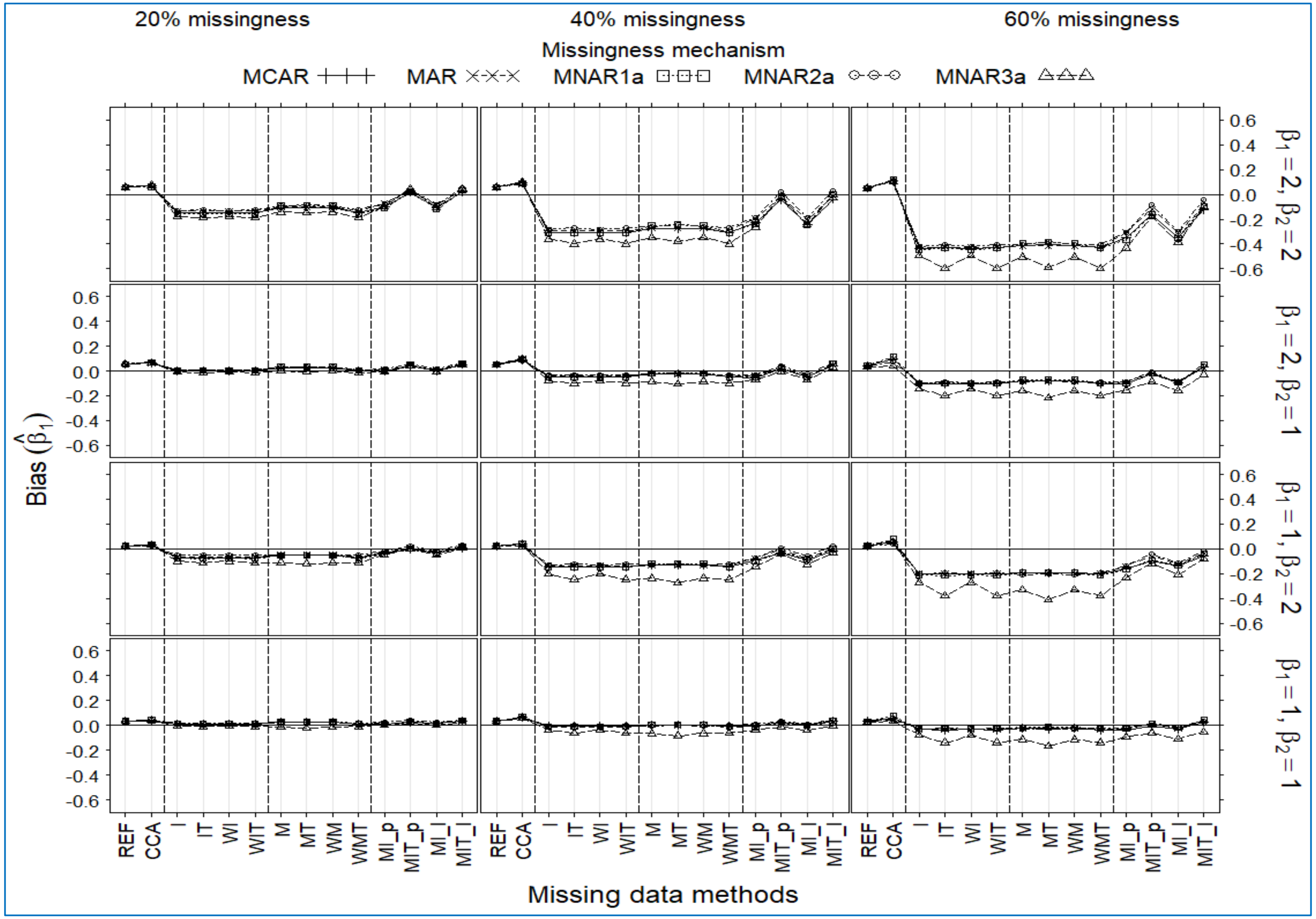
Simulation results for continuous outcome: *Coverage of 95% CI* for $\widehat{\beta}_1$ (Figure 2)



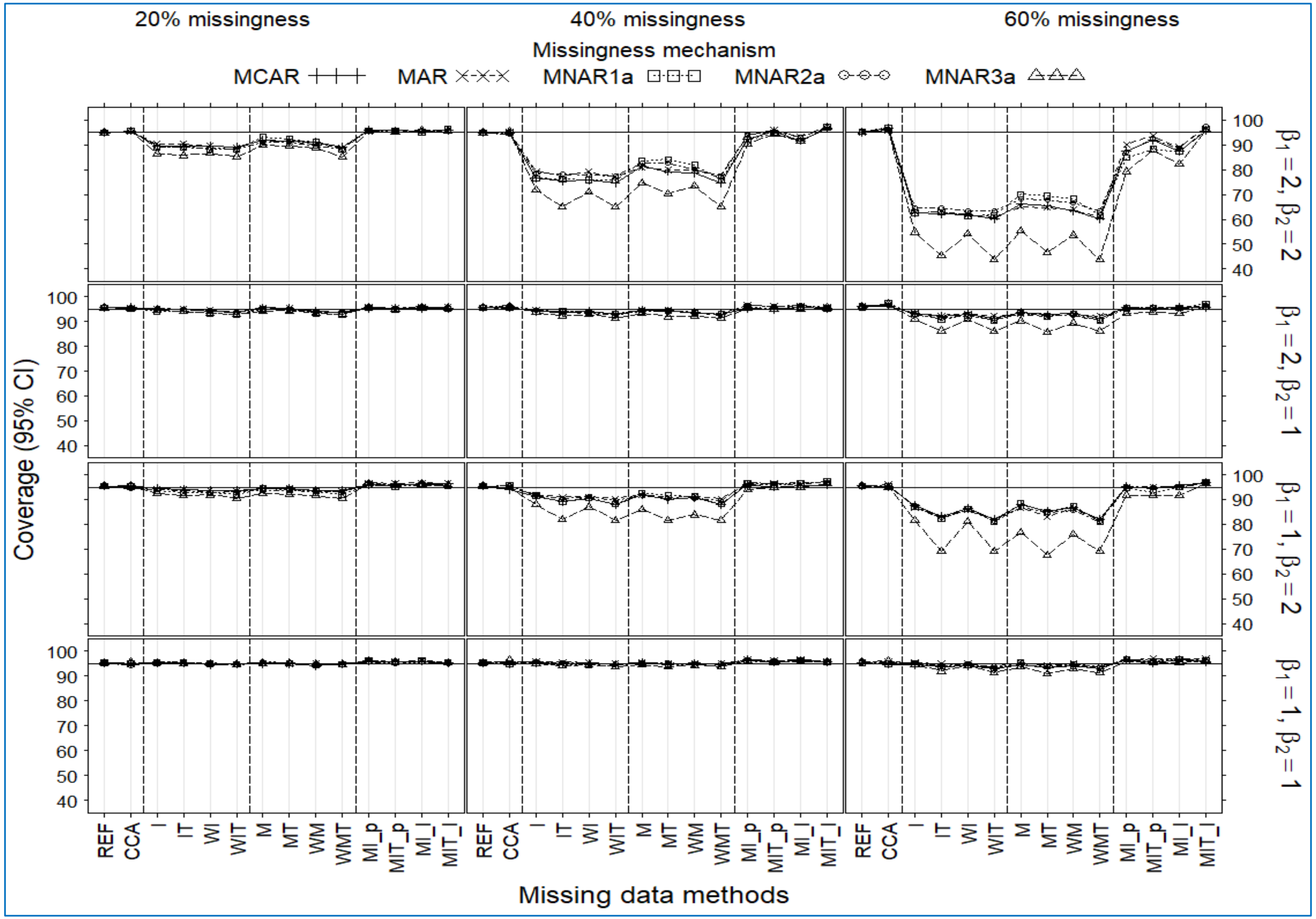
Simulation results for continuous outcome: *RP of $\widehat{\beta}_1$* (Figure 3)



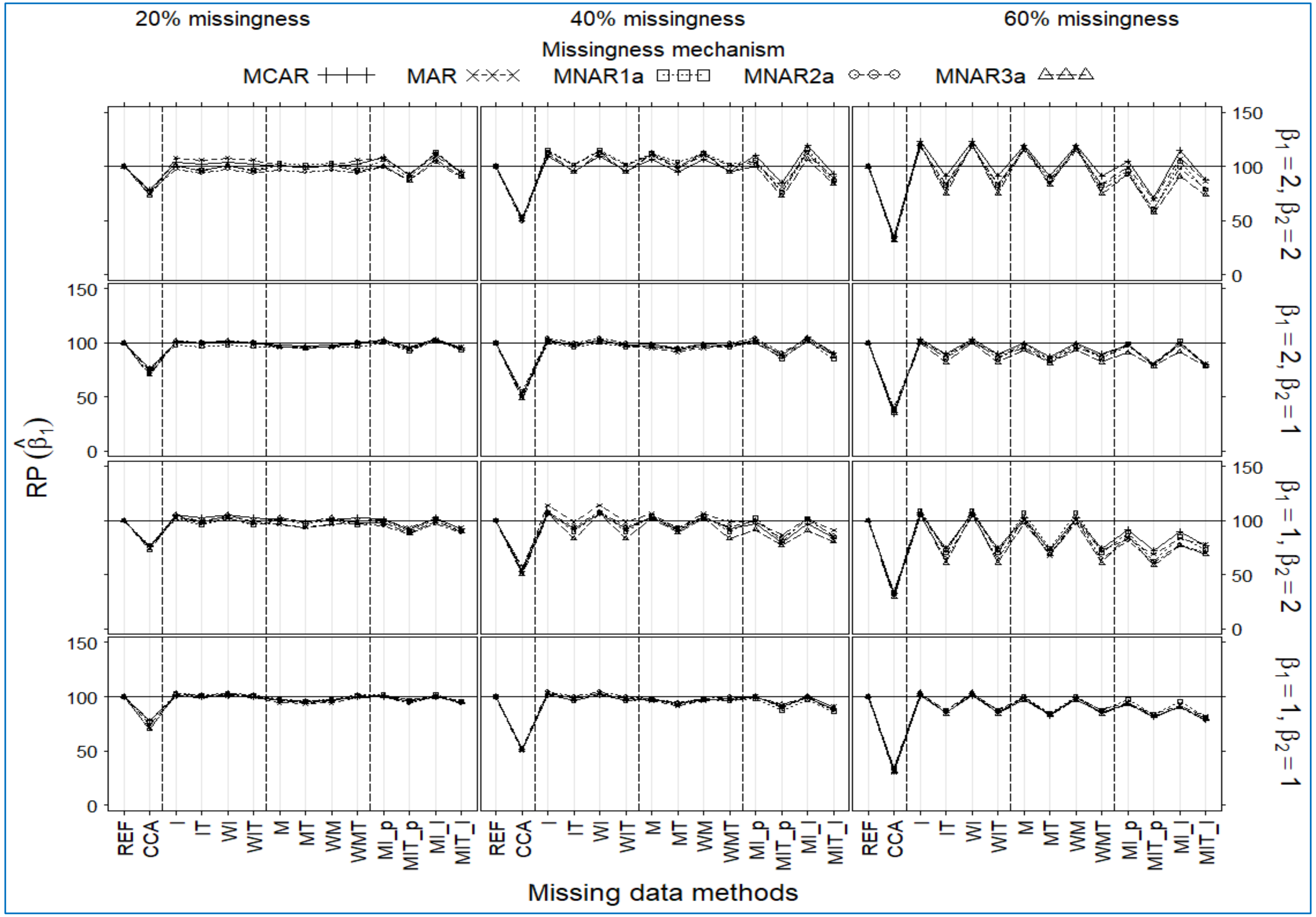
Simulation results for time-to-event outcome: *Bias of $\hat{\beta}_1$* (Figure E1)



Simulation results for time-to-event outcome: *Coverage of 95% CI* for $\hat{\beta}_1$ (Figure E2)



Simulation results for time-to-event outcome: *RP of $\hat{\beta}_1$* (Figure E3)



Discussion for RCT with continuous outcome: *Recommendations (1)*

- No substantial difference in results between the missingness mechanisms, except MNAR3

- Imputation should not be performed per treatment, because this loses precision and underestimates SE, which may result in undercoverage;

- **When missingness is unrelated with treatment:**
  - The missing-indicator method is best;
  - Mean imputation is a good alternative if there is a need to use less covariates in the analysis;
  - MI is not recommended because it is unnecessarily complex for situations similar to ours and always fails to outperform a simple good alternative; and
  - CCA is preferable (easy to perform) only if the proportion of missingness is negligible: so that precision loss is not substantial

Discussion for RCT with continuous outcome: *Recommendations (2)*

➢ **When missingness is related with treatment:**

- It is safe to use mean imputation, since this produces acceptable results across all the applicable missingness mechanisms;
- The missing-indicator method can be used, provided that missingness is not dependent on treatment by covariate interaction: *if it is sure that MNAR3 is implausible*;
- MI is not preferable, for the same reasons provided previously; and
- CCA is preferable only if the proportion of missingness is negligible: *easy to perform and minimal loss of precision*

❑ **Under MNAR3**,

- MI shows some bias probably because *T*Z* was not used in the imputation model;
- The missing-indicator method is seriously biased.

Discussion for RCT with time-to-event outcome: *Recommendations (3)*

- **When missingness is related or not with treatment:**

  - Only CCA and MIT produce unbiased treatment effect estimate, with acceptable coverage;
  - But CCA is substantially less precise even when missingness is low (here 10%);
  - All other methods are biased with substantial undercoverage in several scenarios;

❑ MIT is best and, therefore, recommended for handling missing covariate;

❑ CCA can be used only if the missingness rate is much lower than 10%;

❑ All other methods are not appropriate.

Discussion: *Topics for Future work*

- In RCTs:
  - Situations with missingness in multiple covariates (of mixed types) since these are more likely in practice (*under review*)
    - ✓ For example, a trial with a binary covariate and a continuous outcome measured pre- and post-test, where the covariate and the pre-test outcome are partially missing. This situation allow for comparison of the repeated measurements method with the ANCOVA (used in this study)
  - Situations with joint missingness in covariates and outcome (*under review*)
  - How to improve the missing indicator method in case of MNAR3;
  - How to improve MI in case of MNAR3 (the use of JAV approach?);
- In Cluster randomized trials (CRTs):
  - Situations with joint missingness in covariates and outcome (*under study*)

Thanks for attending!

Questions?