An evaluation of approaches for accommodating interactions and non-linear terms in multiple imputation of incomplete three-level data

Rushani Wijesuriya

Katherine J. Lee, Margarita Moreno-Betancur, John B. Carlin and Anurika P. De Silva

Clinical Epidemiology and Biostatistics Unit

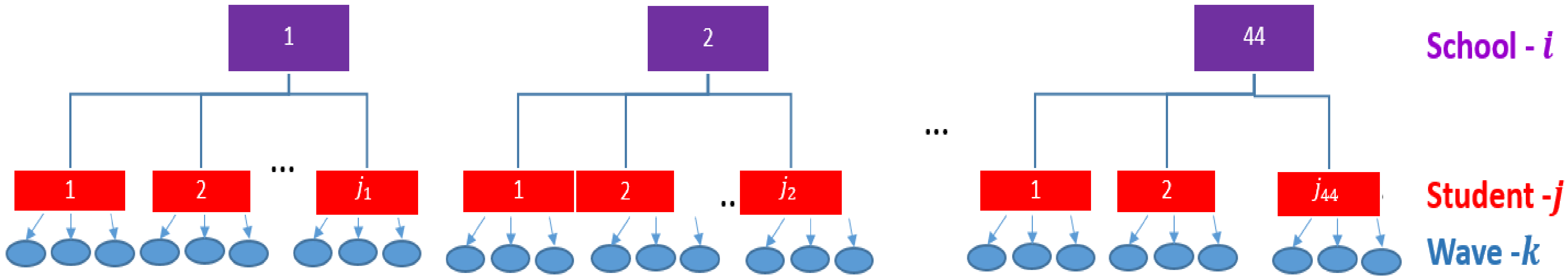
Murdoch Children's Research Institute

The University of Melbourne

04th of November 2020-MiDIA meeting

Background

Childhood to Adolescence Transition Study (CATS) : repeated measures (level 1) of students (level 2) nested within schools (level 3)



In CATS missing data were observed in all of the time-varying variables

The imputation model needed to preserve all the features of the analysis model such as non-linear relationships, interactions and multilevel features[2][3]

Background

Accommodating the three-level structure and interactions or non-linear terms in the imputation model

Accommodating the three-level structure[4]

Extend single-level MI approaches
- School clusters :Dummy indicators (DI)*
- Repeated measures: imputed in wide format

Data configuration

Extend two-level MI approaches
- School clusters : Mixed model based MI
- Repeated measures: imputed in wide format

Data configuration

Extend two-level MI approaches
- School clusters: Dummy indicators (DI)*
- Repeated measures: Mixed model based MI (imputed in long format)

Data configuration

Use three-level MI approaches/Mixed model based MI (repeated measures imputed in long format)

Accommodating interactions or non-linear terms

As repeated measures are in wide format (unless the interaction is with time) ad-hoc extensions will need to be used:
- Impute these terms as just another variable (JAV)
- passively impute these terms after imputation or at each iteration

As the repeated measures are in long format substantive model compatible (SMC) MI can be used

JM-1L-DI-wide
FCS-1L-DI-wide

JM-2L-wide
FCS-2L-wide

SMC-JM-2L-DI
SMC-SM-2L-DI[5]

SMC-JM-3L[6]

*DI extension should be used with caution as it has been shown to produce biased parameter estimates in certain scenarios in some MI literature (7)
*FCS: fully conditional specification, JM: joint modelling, SM: sequential modelling

Aim

Compare MI approaches for imputing incomplete three-level data

- resulting from repeated measures with follow-ups at fixed intervals of time within an individual where there is clustering among individuals (as in the CATS)

- when the substantive analysis model includes interactions or quadratic effects involving incomplete covariates which need to be incorporated in the imputation model

The motivating example :

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

measured using a summary of item scores at waves 2,4 and 6

measured by NAPLAN numeracy scores at waves 3,5 and 7

adjusted for confounders: Child's Sex, SES, NAPLAN scores at wave 1 and Age at wave 1

The Target Analysis Models

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

1. An interaction between the time-varying exposure and time

$$NAPLAN_{ijk} = \beta_0 + \beta_1 \times depression_{ij(k-1)} + \beta_2 \times wave_{ijk} + \beta_3 \times depression_{ij(k-1)} \times wave_{ijk} + ** + b_{oi} + b_{oij} + \varepsilon_{ijk} \quad (1)$$

2. An interaction between the time-varying exposure and a time-fixed baseline variable

$$NAPLAN_{ijk} = \beta_0 + \beta_1 \times depression_{ij(k-1)} + \beta_2 \times wave_{ijk} + \beta_3 \times depression_{ij(k-1)} \times SES_{ij} + ** + b_{oi} + b_{oij} + \varepsilon_{ijk} \quad (2)$$

3. A quadratic effect of the time-varying exposure

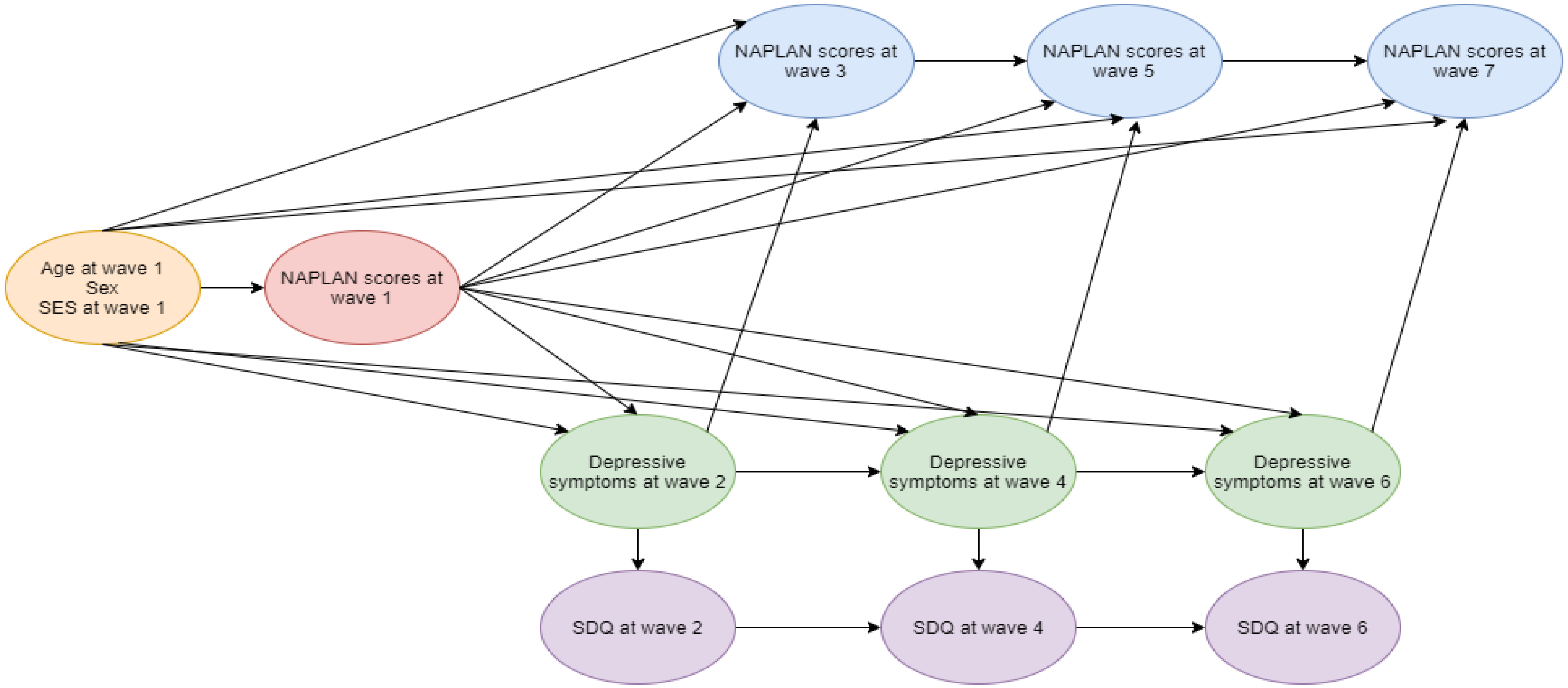
$$NAPLAN_{ijk} = \beta_0 + \beta_1 \times depression_{ij(k-1)} + \beta_2 \times wave_{ijk} + \beta_3 \times depression^2_{ij(k-1)} + ** + b_{oi} + b_{oij} + \varepsilon_{ijk} \quad (3)$$

With $b_{oi} \sim N\left(0, \sigma^2_{boi}\right), b_{oij} \sim N\left(0, \sigma^2_{boij}\right) and\ \varepsilon_{ijk} \sim N(0, \sigma^2_{\varepsilon ijk})$

**the remaining covariates that were adjusted for in (1),(2) and (3) include Child's Sex, SES, NAPLAN scores at wave 1 and Age at wave 1

Simulation Study

The data were generated by mimicking the CATS data which was replicated 1000 times



We also considered two different numbers of higher level clusters: 40 school clusters and 10 school clusters

Missing values generated

- exposure (15%, 20% and 30% of the depressive symptom scores at waves 2,4, and 6 respectively) according to a MAR mechanism
- a time-fixed confounder (10 % of Socio-Economic Status) according to a MCAR mechanism.

MI Approaches

	How the two sources of clustering are handled		How the approach accommodate interactions/non-linear terms		
MI approach	**Clustering due to higher level clusters**	**Clustering due to repeated measures**	**Interaction between the time-varying exposure and time**	**Interaction between the time-varying exposure and a time-fixed baseline variable**	**Quadratic effect of the exposure**
JM-1L-DI-wide	DI	Repeated measures imputed in wide format	Repeated measures imputed in wide format	Not accommodated (ad-hoc extensions can be used but are not congenial with substantive analysis)	Not accommodated (ad-hoc extensions can be used but are not congenial with substantive analysis)
FCS-1L-DI-wide	DI	Repeated measures imputed in wide format	Repeated measures imputed in wide format		
JM-2L-wide	RE	Repeated measures imputed in wide format	Repeated measures imputed in wide format		
FCS-2L-wide	RE	Repeated measures imputed in wide format	Repeated measures imputed in wide format		
SMC-JM-2L-DI	DI	RE	Through SMC-MI algorithm+	Through SMC-MI algorithm +	Through SMC-MI algorithm +
SMC-SM-2L-DI	DI	RE	Through SMC-MI algorithm +	Through SMC-MI algorithm +	Through SMC-MI algorithm +
SMC-JM-3L	RE	RE	Through SMC-MI algorithm ++	Through SMC-MI algorithm ++	Through SMC-MI algorithm ++

MI Approaches

Analysis model (1)

1. ***JM-1L-DI-wide***
2. ***FCS-1L-DI-wide***
3. ***JM-2L-wide***
4. ***FCS-2L-wide***
5. ***SMC-JM-2L-DI***
6. ***SMC-SM-2L-DI***
7. ***SMC-JM-3L***

Analysis model (2)

JM : JAV to incorporate the interaction

1. ***JM-1L-DI-wide-JAV***
2. ***JM-2L-wide-JAV***

FCS : passive imputation within iterations using two variations of reverse imputation strategy[(8),(9)]

3. ***FCS-1L-DI-wide-passive_c***
4. ***FCS-2L-wide-passive _c***
5. ***FCS-1L-DI-wide-passive_all***
6. ***FCS-2L-wide-passive_all***

7. ***SMC-JM-2L-DI***
8. ***SMC-SM-2L-DI***
9. ***SMC-JM-3L***

For benchmark

10. ***JM-1L-DI-wide***
11. ***FCS-1L-DI-wide***

Analysis model (3)

1. ***JM-1L-DI-wide-JAV***
2. ***JM-2L-wide-JAV***
3. ***FCS-1L-DI-wide-passive***
4. ***FCS-2L-wide-passive***
5. ***SMC-JM-2L-DI***
6. ***SMC-SM-2L-DI***
7. ***SMC-JM-3L***

For benchmark

8. ***JM-1L-DI-wide***
9. ***FCS-1L-DI-wide***

Passive reverse imputation strategy

- **passive concurrent (passive_c)**

Imputing depressive symptom values at a particular wave:
Single interaction between the NAPLAN score at the next wave and SES as a predictor

Depressive symptoms at wave 2 → **Interaction between SES and NAPLAN at wave 3**

Depressive symptoms at wave 4 → **Interaction between SES and NAPLAN at wave 5**

Depressive symptoms at wave 6 → **Interaction between SES and NAPLAN at wave 7**

Imputing SES:
Interactions between the NAPLAN scores and depressive symptom scores at previous wave for all 3 waves as predictors

SES → **Interaction between depressive symptoms at wave 2 and NAPLAN at wave 3**
Interaction between depressive symptoms at wave 4 and NAPLAN at wave 5
Interaction between depressive symptoms at wave 6 and NAPLAN at wave 7

To allow the association between the outcome and exposure at each wave to vary for different levels of SES and vice versa as implied by the substantive analysis model

Passive reverse imputation strategy

- **passive all (passive_all)**

Imputing depressive symptom values at a particular wave:
Interactions between the NAPLAN scores at each of the 3 waves and SES as predictors

Depressive symptoms at wave 2 → **Interaction between SES and NAPLAN at wave 3**
Interaction between SES and NAPLAN at wave 5
Interaction between SES and NAPLAN at wave 7

Same for depressive symptoms at wave 4, and 6

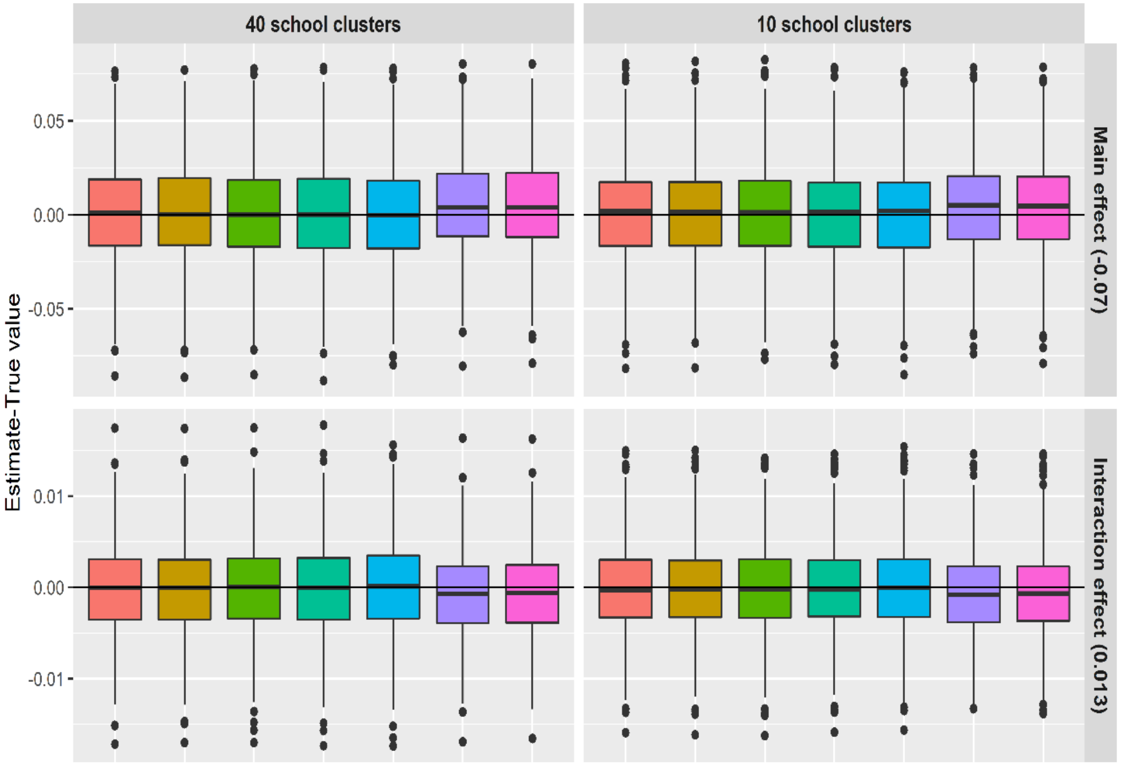
Imputing SES:
Interactions between the NAPLAN scores and depressive symptom scores at previous wave for all 3 waves as predictors

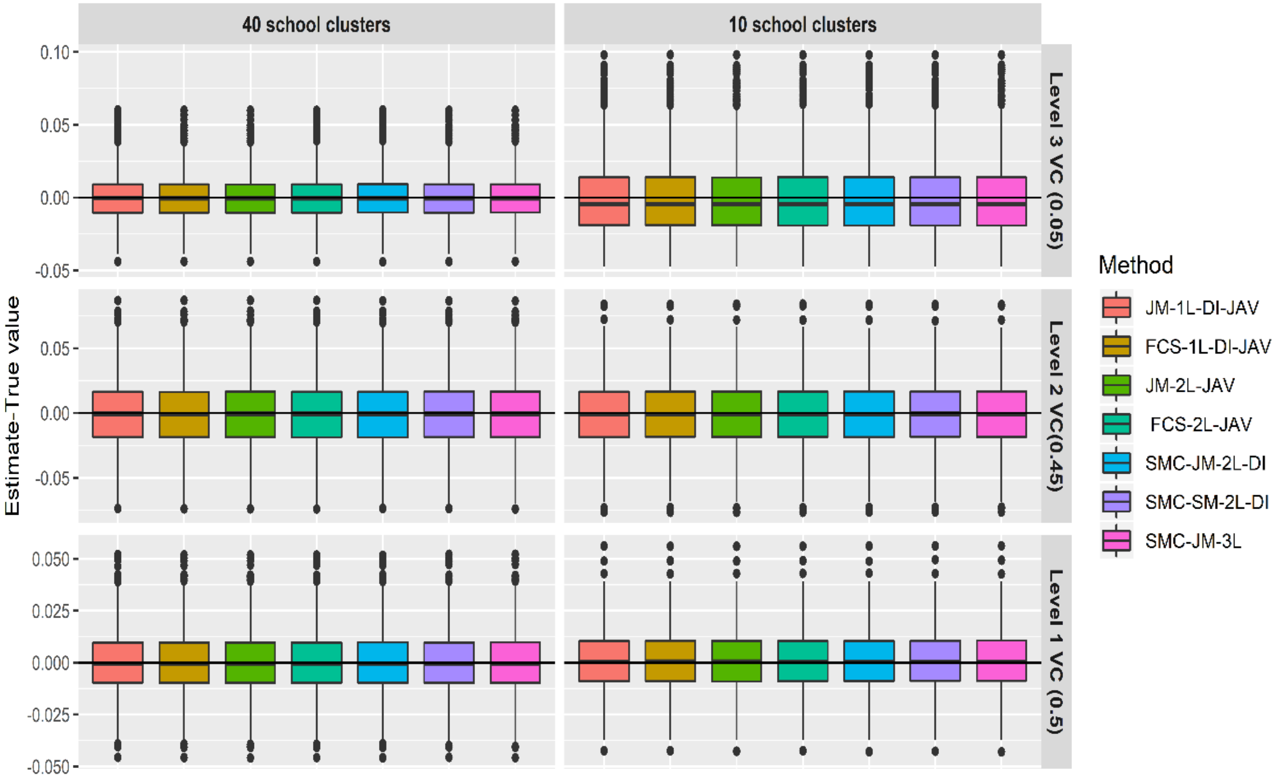
SES → **Interaction between depressive symptoms at wave 2 and NAPLAN at wave 3**
Interaction between depressive symptoms at wave 4 and NAPLAN at wave 5
Interaction between depressive symptoms at wave 6 and NAPLAN at wave 7

Allows the association between the outcome and the exposure to vary for different levels of SES and vice versa, but allows even more flexibility

Results (Bias)-Analysis Model 1

Interaction between the time-varying exposure and time

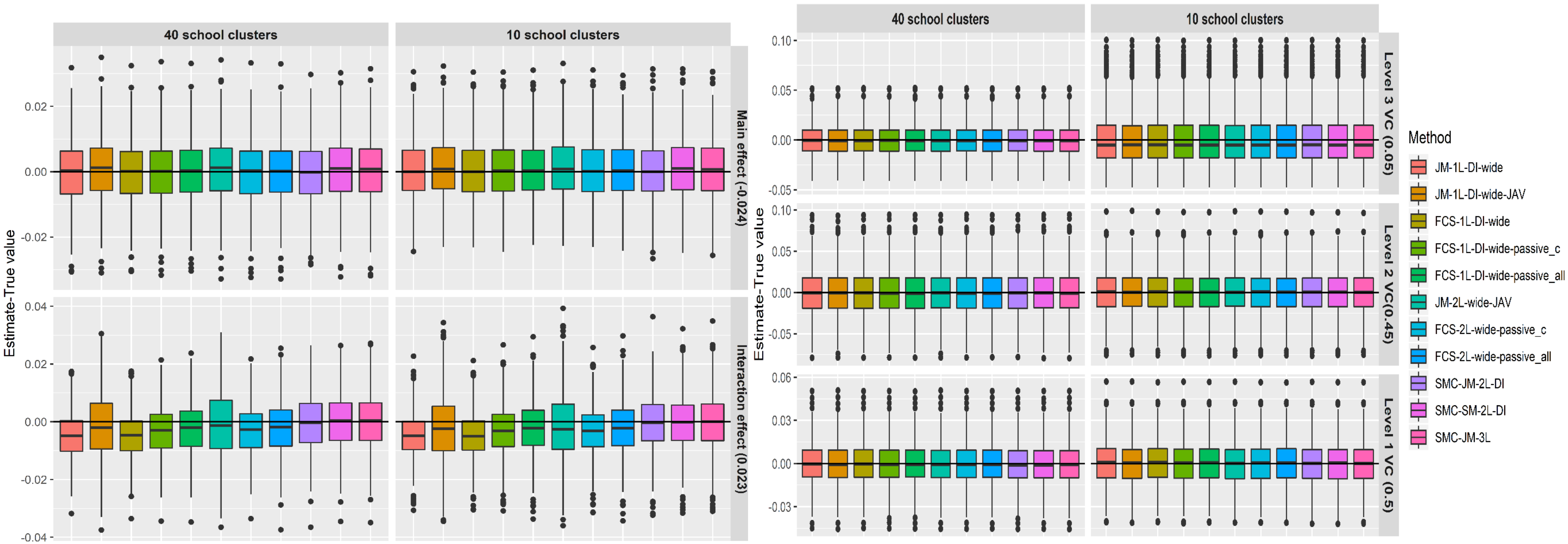




- All the MI approaches produced approximately unbiased estimates of the main effect and the interaction effect
- All approaches resulted in similar negligible bias (<10% relative bias) for the 3 variance components

Results (Bias)-Analysis Model 2

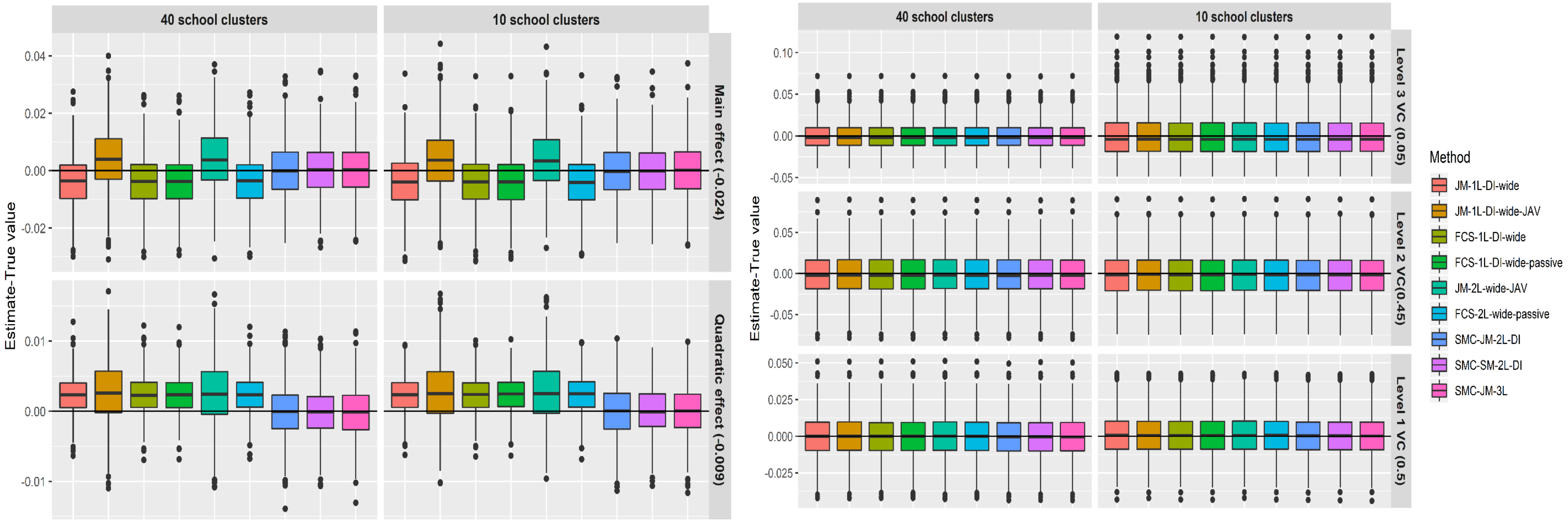
Interaction between the time-varying exposure and a time-fixed baseline variable



- All of the MI approaches except for ***SMC-JM-2L-DI, SMC-SM-2L-DI*** and ***SMC-JM-3L*** resulted in biased estimates for the interaction effect, with substantial underestimation of the interaction effect
- All approaches resulted in similar negligible bias (<10% relative bias) for the 3 variance components

Results (Bias)-Analysis Model 3

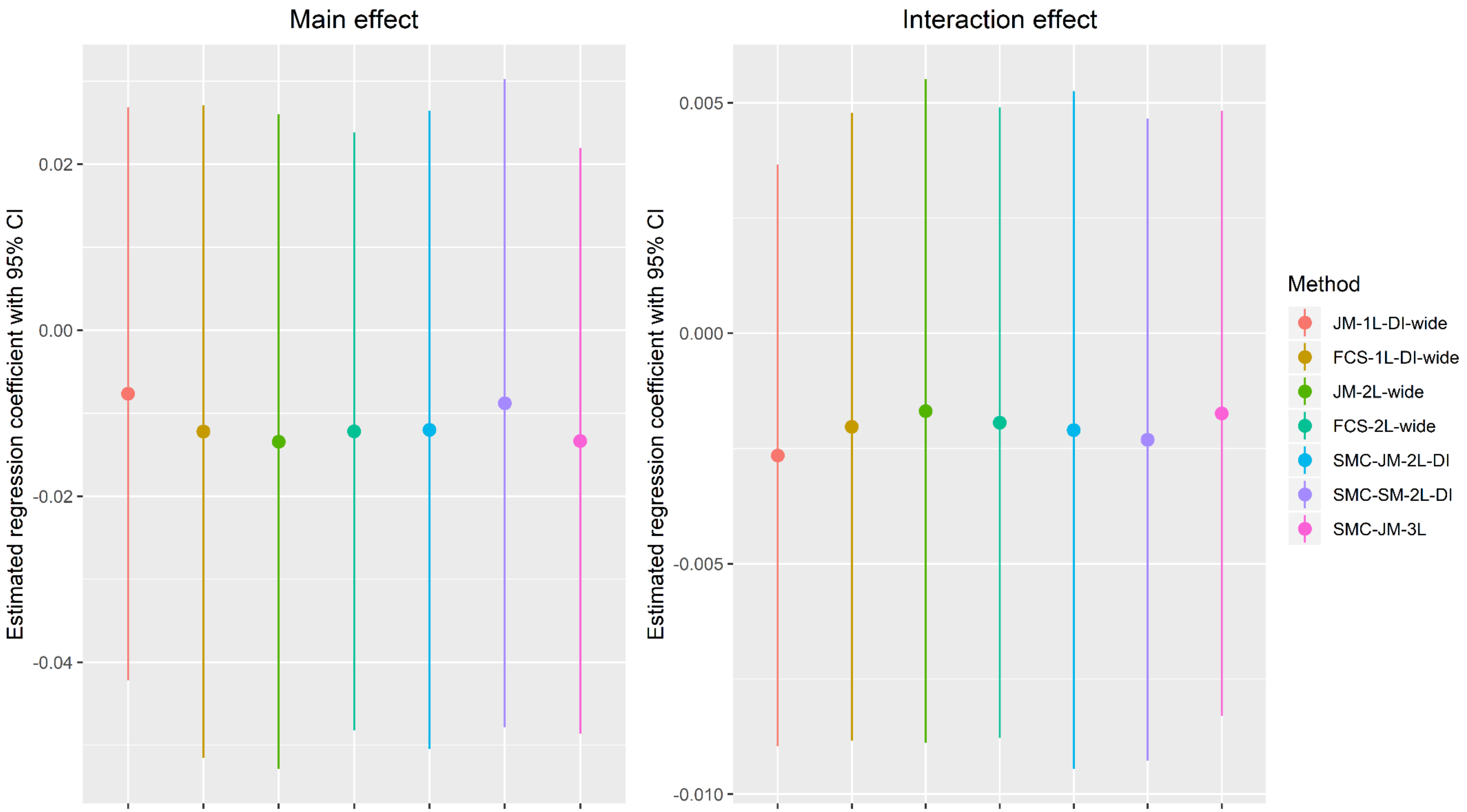
Quadratic effect of the time-varying exposure



- All of the MI approaches except for ***SMC-JM-2L-DI, SMC-SM-2L-DI*** and ***SMC-JM-3L*** resulted in biased estimates for the quadratic term, with substantial underestimation of the quadratic effect
- All approaches resulted in similar negligible bias (<10% relative bias) for the variance components

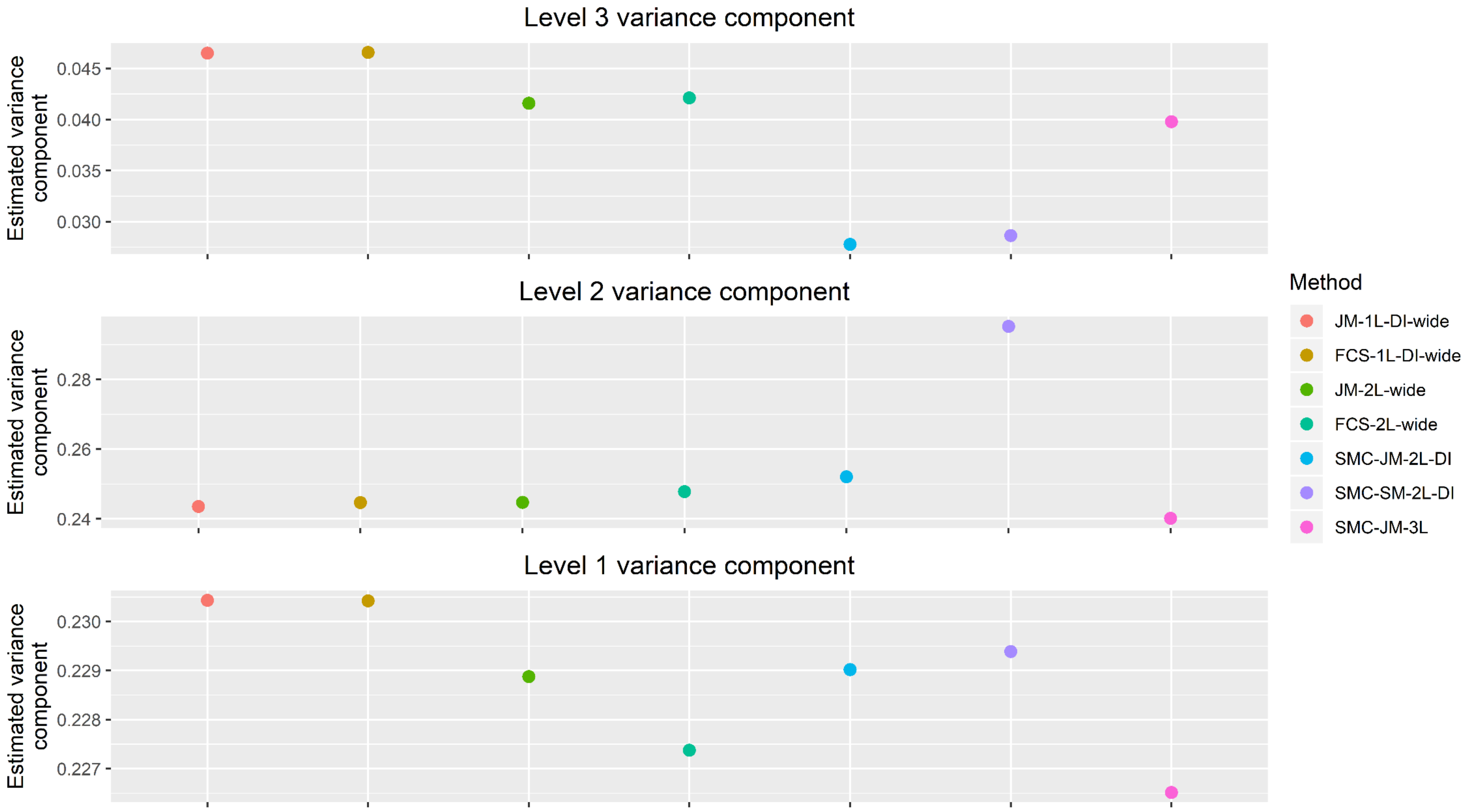
CATS application-Analysis Model 1

Interaction between the time-varying exposure and time



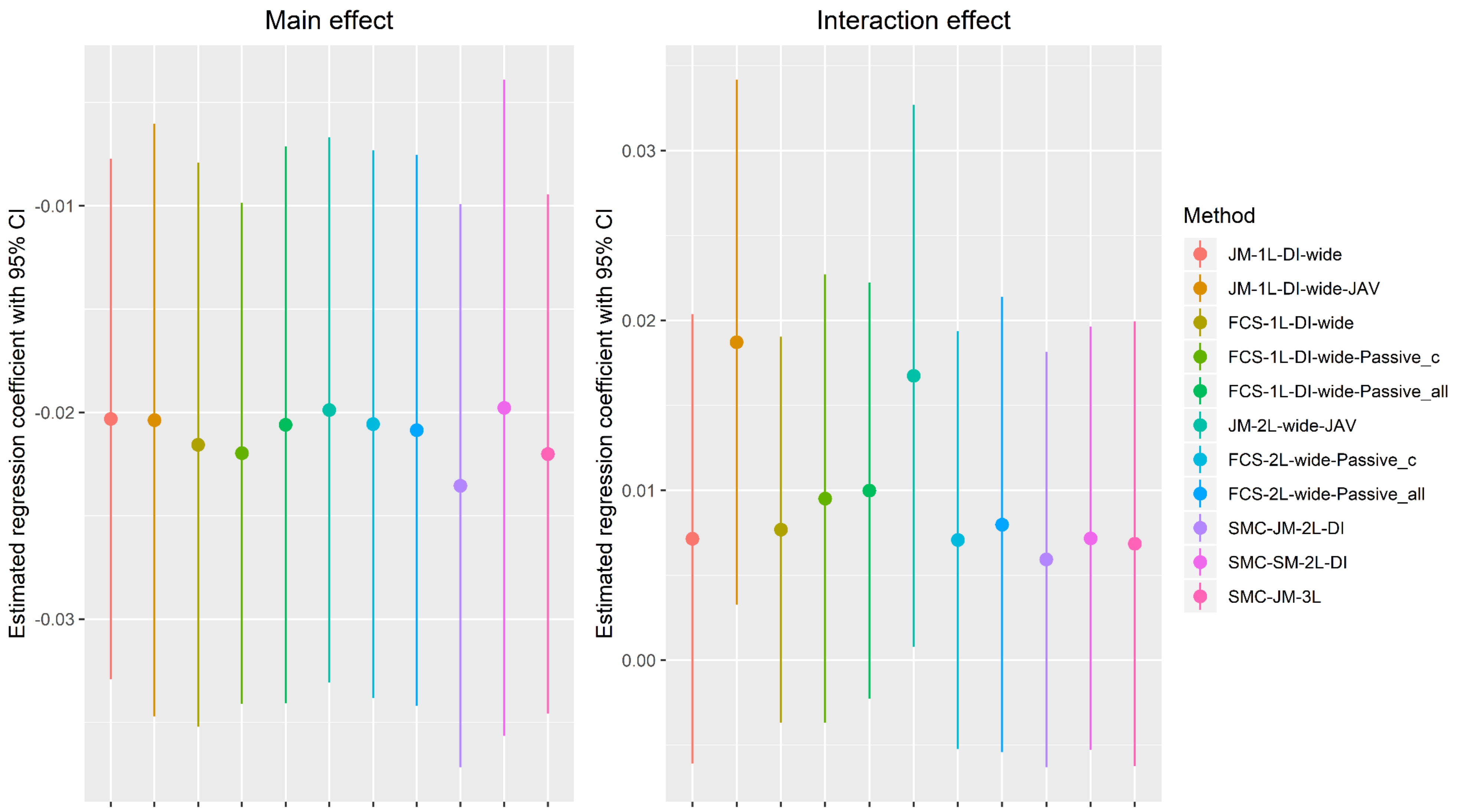
CATS application-Analysis Model 1

Interaction between the time-varying exposure and time



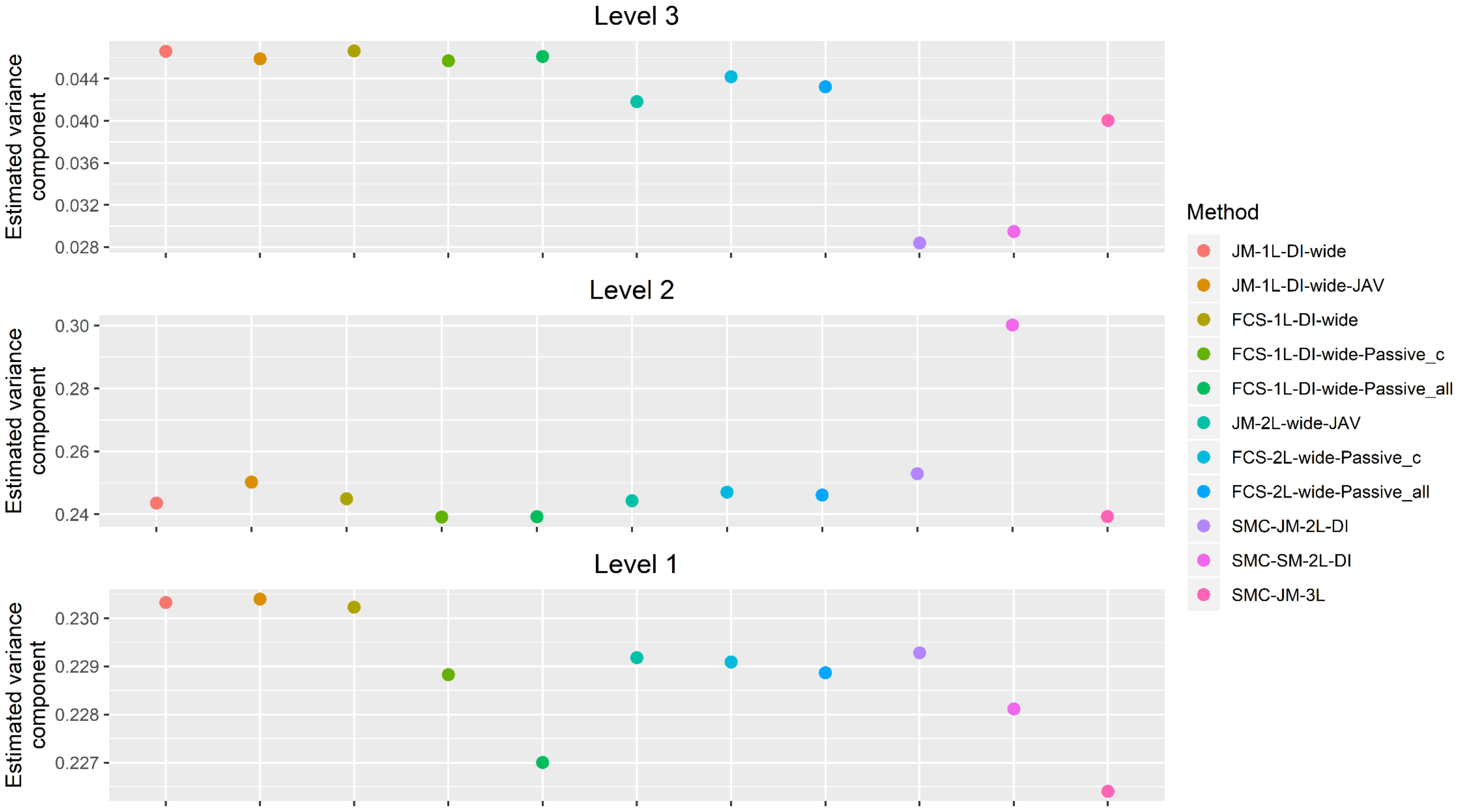
CATS application-Analysis Model 2

Interaction between the time-varying exposure and a time-fixed baseline variable



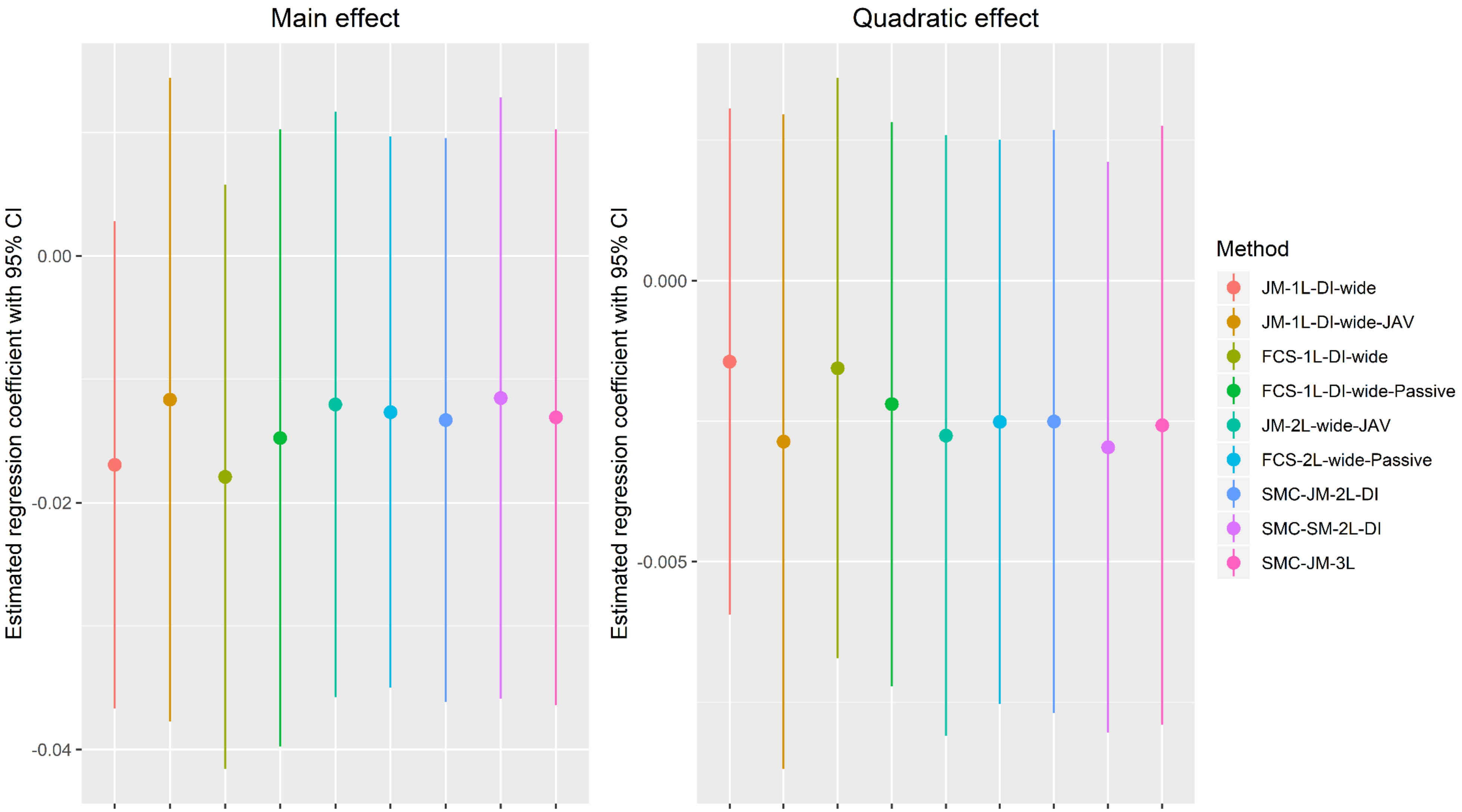
CATS application-Analysis Model 2

Interaction between the time-varying exposure and a time-fixed baseline variable



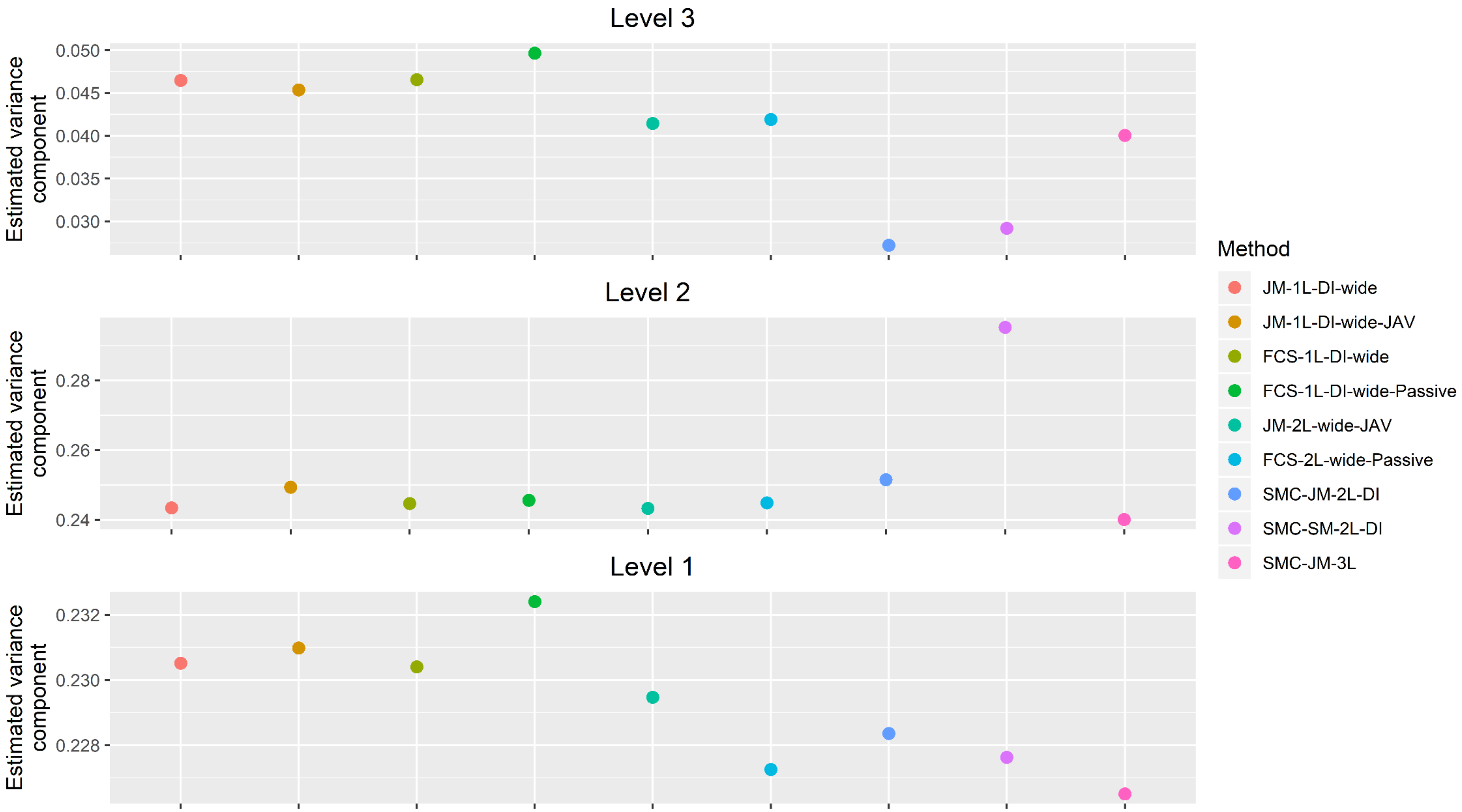
CATS application-Analysis Model 3

Quadratic effect of the time-varying exposure



CATS application-Analysis Model 3

Quadratic effect of the time-varying exposure



Conclusions

- With an analysis model where there is an interaction with time, all of the approaches (including the single-level and two-level adaptations) considered seem to be appropriate
- However, approaches which use the DI extension should be used with caution as they can be problematic in certain scenarios(7)
- When the analysis model involves an interaction between the time-varying exposure and an incomplete time-fixed confounder or quadratic effects the three-level SMC approach is recommended

References

(1) Rezvan, P. H., Lee, K. J. & Simpson, J. A. 2015. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC med res methodology*, 15, 30.
(2) Meng, X.-L. 1994. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science***,** 538-558
(3) Bartlett, J. W., Seaman, S. R., White, I. R. & Carpenter, J. R. 2015. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical methods in medical research,* 24**,** 462-487
(4) Wijesuriya, R. *et al.* Evaluation of approaches for multiple imputation of three-level data. *BMC Med Res Methodology* **20,** 207 (2020).
(5) Lüdtke, O., Robitzsch, A. & West, S. G. 2019. Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological methods*
(6) Enders, C. K., Du, H. & Keller, B. T. 2019. A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological methods*
(7) Drechsler J. Multiple imputation of multilevel missing data—rigor versus simplicity. J Educ Behav Stat. 2015;40(1):69–95.
(8) Grund S, Lüdtke O, Robitzsch A. Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods.* 2018;21(1):111-149.
(9) van Buuren S. *Flexible imputation of missing data.* Chapman and Hall/CRC; 2018

Additional Resources

- Software code written for the simulation studies can be found here https://github.com/rushwije/MI_three-level
- Our previous paper evaluating MI approaches for incomplete three-level data: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01079-8
- A pre-print of this work :

  http://arxiv.org/abs/2010.16025

Extra Slides

Data configurations

Single-level MI approaches with DI indicators and repeated measures imputed in wide format

With one row per individual

ID	Age	SES	Prev_dep.3	Prev_dep.5	Prev_dep.7	School cluster
11011	9.949622	1032.9673	0	0	0	19
11021	9.087175	1070.199	1	1	2	19
11031	8.287702	1070.199	0	.	.	22
11041	8.884569	1040.2396	.	2	8	20
11051	9.574527	1070.199	4	1	2	21
11061	8.821597	1009.2175	2	.	0	18
11071	9.248713	1070.199	1	0	0	19

Include as DIs or as a categorical variable in single-level the imputation model

Overview of MI approaches

Two -level MI approach with repeated measures imputed in wide format

With one row per individual and repeated measures in wide format

ID	Age	SES	Prev_dep.3	Prev_dep.5	Prev_dep.7	School cluster
11011	9.949622	1032.9673	0	0	0	19
11021	9.087175	1070.199	1	1	2	19
11031	8.287702	1070.199	0	.	.	22
11041	8.884569	1040.2396	.	2	8	20
11051	9.574527	1070.199	4	1	2	21
11061	8.821597	1009.2175	2	.	0	18
11071	9.248713	1070.199	1	0	0	19

Include as cluster indicator /group variable in the two-level imputation model

Overview of MI approaches

Two -level MI approach with DI approach

With one row per individual per wave (long format)

Include as cluster indicator /group variable in the two-level imputation model

ID	Age	SES	Prev_dep	Wave	School cluster
11011	9.949622	1032.9673	0	3	19
11011	9.949622	1032.9673	0	5	19
11011	9.949622	1032.9673	0	7	19
11021	9.087175	1070.199	1	3	19
11021	9.087175	1070.199	1	5	19
11021	9.087175	1070.199	2	7	19
11031	8.287702	1070.199	0	3	22
11031	8.287702	1070.199	.	5	22
11031	8.287702	1070.199	.	7	22

Include as DIs or as a categorical variable in the two-level imputation model

[Overview of MI approaches]