Multiple Imputation with Missing Indicators as Proxies for Unmeasured Variables: Simulation Study

Matthew Sperrin (matthew.sperrin@manchester.ac.uk)
The University of Manchester https://orcid.org/0000-0002-5351-9960
Glen P. Martin
The University of Manchester

Research article

Keywords: Missing data, Missing indicator, Multiple imputation, Simulation study

Posted Date: June 26th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-24268/v3

Version of Record: A version of this preprint was published on July 8th, 2020. See the published version at https://doi.org/10.1186/s12874-020-01068-x.

Multiple Imputation with Missing Indicators as Proxies for Unmeasured Variables: Simulation Study

Matthew Sperrin[*1] and Glen P. Martin[1]

[1]: Faculty of Biology, Medicine and Health, University of Manchester.

[*]: Correspondence to: Matthew Sperrin, Vaughan House, University of Manchester, Manchester, M13 9PL

matthew.sperrin@manchester.ac.uk

Running headline: Missing Indicators and Unmeasured Variables

Abstract

Background: Within routinely collected health data, missing data for an individual might provide useful information in itself. This occurs, for example, in the case of electronic health records, where the presence or absence of data is informative. While the naive use of missing indicators to try to exploit such information can introduce bias, its use in conjunction with multiple imputation may unlock the potential value of missingness to reduce bias in causal effect estimation, particularly in missing not at random scenarios and where missingness might be associated with unmeasured confounders.

Methods: We conducted a simulation study to determine when the use of a missing indicator, combined with multiple imputation, would reduce bias for causal effect estimation, under a range of scenarios including unmeasured variables, missing not at random, and missing at random mechanisms. We use directed acyclic graphs and structural models to elucidate a variety of causal structures of interest. We handled missing data using complete case analysis, and multiple imputation with and without missing indicator terms.

Results: We find that multiple imputation combined with a missing indicator gives minimal bias for causal effect estimation in most scenarios. In particular the approach: 1) does not introduce bias in missing (completely) at random scenarios; 2)reduces bias in missing not at random scenarios where the missing mechanism depends on the missing variable itself; and 3) may reduce or increase bias when unmeasured confounding is present.

Conclusion: In the presence of missing data, careful use of missing indicators, combined with multiple imputation, can improve causal effect estimation when missingness is informative, and is not detrimental when missingness is at random.

Keywords: Missing data; Missing indicator; Multiple imputation; Simulation study

Background

Missing data is a common feature in observational studies. It is conventional to view missing data as a nuisance, and as such, methods to handle missing data usually target an estimand that would be available in the absence of missing data (completed data estimand). The mechanism for missingness is conventionally divided into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [1, 2]. In the case of MCAR and MAR, an unbiased estimator of any completed data estimand exists. Such an estimator is provided by complete case analysis in MCAR scenarios, and by multiple imputation in both MAR and MCAR scenarios. In contrast, under MNAR, unbiased estimators of a given completed data estimand may or may not exist, depending on the nature of the estimand, and the joint distribution of the missingness mechanism and the other variables under consideration [2].

Alongside missing data, confounding is a threat to causal effect estimation in observational studies, especially where this is caused by unmeasured variables. Where unmeasured confounding exists, it is not possible to construct unbiased estimators of a causal effect, without making strong, unverifiable assumptions [3].

For example, consider a scenario in which we are interested in calculating the causal effect of total cholesterol (exposure) on cardiovascular disease (outcome), using electronic health records. Presence (analogous to missingness) of a cholesterol test result for a particular patient indicates that a decision was made to run this test, and the reason for this decision is likely to depend on characteristics of the patient; for example the patient's diet, which may or may not be recorded. Diet may affect both the result of the laboratory test, and the

outcome of interest, hence confounding. If information concerning diet is not recorded, we therefore have unmeasured confounding, and unbiased estimators of the causal estimand may not exist. Moreover, the missingness mechanism for the exposure may depend on unmeasured variables, in which case the exposure is MNAR, and an unbiased estimator of the completed data estimand may not exist either.

An emerging hypothesis is that in scenarios such as this, missing data may be a blessing rather than a curse, because the missingness mechanism can be used as a proxy for the unmeasured confounding, through the use of missing indicators [4]. Suppose one wished to use missing data approaches to target the causal estimand directly (rather than the completed data estimand, as is done conventionally [5]). Then, exploiting the missingness mechanism through the use of missing indicators may reduce bias even compared with estimation in the absence of missing data, particularly when the missing indicator is used in conjunction with multiple imputation (MIMI) [4, 6–8]. This is despite that naïve use of missing indicators introduces bias in the completed data estimand under MAR and MCAR [9–12].

In this paper we investigate, through simulation supplemented with analytical findings, the potential for using the missingness mechanism to partly adjust for unmeasured confounding and other missing not at random scenarios, and identify the cases where this can reduce bias for causal effect estimation.

Methods

Scenarios and data generating mechanisms

Our aim is to identify missing data strategies that recover causal effects of an exposure on an outcome, with minimal bias in a variety of scenarios, especially where the causal effects are affected by unmeasured confounding. The scenarios that we consider in this paper are given in Figure 1. We consider a partially observed exposure A, a fully observed outcome Y and a further variable U, which is either fully unobserved or fully observed depending on the mechanism for missingness. The missingness indicator for A is R_A where $R_A = 0$ when A is missing. In the example in the Introduction, A is total cholesterol, Y is cardiovascular disease, U is diet, and R_A denotes whether a cholesterol test has been performed or not. A^* is the observable part of A, i.e. $A^* = A$ when $R_A = 1$, and missing when $R_A = 0$. So A^* is what we observe, while A includes unobserved values.

We use the counterfactual notation for consideration of causal effects, e.g. $Y(A = a)$ denotes the value of Y that would be observed if, possibly contrary to fact, we set $A = a$, and we will abbreviate to $Y(a)$ where this does not lead to ambiguity. See [3] for an introduction to causal inference with counterfactuals. Our primary aim is to recover the unconditional causal effect of A on Y; for continuous exposure, A, that is the expected effect on Y for 1-unit increase in A: $\delta_A := E[Y(A = a + 1) - Y(A = a)]$. We also have a secondary interest in inferring the presence of unmeasured confounding (i.e. whether an unobserved U directly affects both A and Y) or missing not at random mechanisms, or both.

First, we consider scenarios in Figure 1 where U is assumed to be unobserved, which we label (i)-(vi). Scenario (i) corresponds to MCAR, since R_A is independent of all other variables. All other scenarios, (ii)-(vi), are MNAR since R_A is dependent on U or A or both. Note here that we follow the graphical definitions of MCAR, MAR and MNAR as set out in [2]. In scenarios (i) and (iv), complete case analysis can yield unbiased estimates of the causal effect of A on Y (see e.g. [13] for scenario (iv)). In scenarios (iii) and (vi), the unobserved variable U confounds the relationship between A and Y, so an unbiased estimate of the causal effect of A on Y may not be available even if there were no missingness.

In scenarios (ii), (iii), (v), and (vi), we could view R_A as a proxy for the unobserved U. It therefore may be beneficial to include R_A in the outcome model. This may reduce bias in the estimation of the causal effect of A on Y, by partly adjusting for the confounding effect of U.

Second, we consider each of the six scenarios in Figure 1 with U assumed fully observed, and label these (i-U)-(vi-U). Here, we do not expect any benefit in including R_A in the outcome model, but we wish to examine any reduction in performance that doing so may introduce. Scenario (i-U) remains MCAR, while scenarios (ii-U) and (iii-U) are now MAR. Scenarios (iv-U), (v-U), and (vi-U) remain MNAR but only through the dependence of R_A on A.

We now specify the structural models that will be assumed for our further derivations and simulations.

- U is binary with $P[U = 1] = \pi_U$.

- A is continuous, with $A \sim N(\alpha_0 + \alpha_U U, \sigma_A^2)$.
- R_A is binary, with either $P[R_A = 0] = \text{expit}(\beta_0 + \beta_U U + \beta_A A + \beta_{UA} UA)$, or simply $R_A = 1 - U$, depending on the scenario considered.
- Y is continuous, with $Y \sim N(\gamma_0 + \gamma_U U + \gamma_A A + \gamma_{UA} UA, \sigma_Y^2)$.

The outcome model is linear in A hence the causal effect of a one-unit change in A does not depend on the starting value of A, however it does depend on U because of the interaction term. Specifically, the (true) unconditional causal effect of interest, by standardization, is $\delta_A = E[Y(A = a + 1) - Y(A = a)] = \gamma_A + \pi_U \gamma_{UA}$.

Considered approaches

For notation, we use Greek letters with no superscripts to denote true parameter values (from the data generating mechanisms described in the previous section) – e.g. γ_A - and use the same Greek letters with bracketed superscripts to denote the parameters estimated in the various analysis models – e.g. $\gamma_A^{(1)}$. We consider the following imputation and modelling approaches.

First, a complete case analysis. When U is unobserved, this fits the model $E[Y] = \gamma_0^{(0)} + \gamma_A^{(0)} A^*$, restricting to observations where $R_A = 1$. When U is observed, the model is $E[Y] = \gamma_0^{(0U)} + \gamma_A^{(0U)} A^* + \gamma_U^{(0U)} U + \gamma_{UA}^{(0U)} UA^*$.

Second, we consider multiple imputation, under a joint normal model assuming a MAR mechanism. Thus, when U is unobserved the imputation model is $E[A] = \phi_0^{(I)} + \phi_Y^{(I)} Y$, and

when U is observed the imputation model is $E[A] = \phi_0^{(IU)} + \phi_Y^{(IU)}Y + \phi_U^{(IU)}U + \phi_{UY}^{(IU)}UY$ (including the interaction term, as recommended in [14]). Missing As are imputed by a random draw from the predictive distribution implied by the imputation model, using 'proper' imputation which accounts for both the uncertainty in the imputation model, and the residual variance[15]. This is repeated multiple times and subsequent results are pooled over iterations using Rubin's rules (in this study we consider five imputations for the sake of computational time).

Throughout, we denote the imputed A as A_{imp}. We then consider the following three outcome/analysis models, when U is unobserved:

1. 'MI(A)': $E[Y] = \gamma_0^{(1)} + \gamma_A^{(1)}A_{\text{imp}}$.

2. 'MI(R+A)': $E[Y] = \gamma_0^{(2)} + \gamma_A^{(2)}A_{\text{imp}} + \gamma_R^{(2)}(1 - R_A)$.

3. 'MI(R*A)': $E[Y] = \gamma_0^{(3)} + \gamma_A^{(3)}A_{\text{imp}} + \gamma_R^{(3)}R_A + \gamma_{RA}^{(3)}A_{\text{imp}}(1 - R_A)$.

Model 1 represents a standard multiple imputation (MI) approach, while models 2 and 3 are variants of the MIMI approach, without and with interaction (MI(R+A), MI(R*A)).

When U is observed we consider three outcome models of the same form:

1U. 'MI(A)': $E[Y] = \gamma_0^{(1U)} + \gamma_A^{(1U)}A_{\text{imp}} + \gamma_U^{(1U)}U + + \gamma_{UA}^{(1U)}UA_{\text{imp}}$.

2U. 'MI(R+A)': $E[Y] = \gamma_0^{(2U)} + \gamma_A^{(2U)}A_{\text{imp}} + \gamma_R^{(2U)}(1 - R_A) + \gamma_U^{(2U)}U + + \gamma_{UA}^{(2U)}UA_{\text{imp}}$.

3U. 'MI(R*A)': $E[Y] = \gamma_0^{(3U)} + \gamma_A^{(3U)}A_{\text{imp}} + \gamma_R^{(3U)}R_A + \gamma_{RA}^{(3U)}A_{\text{imp}}(1 - R_A) + \gamma_U^{(3U)}U + + \gamma_{UA}^{(3U)}UA_{\text{imp}}$.

Finally, we also include 'completed data' models in which we use the original variable A in our models. This serves as a ground-truth for all analyses in the absence of missing data. In scenarios (i)-(vi), U is not observed hence the completed data model is $E[Y] = \gamma_0^{(C)} + \gamma_A^{(C)}A$, while in scenarios (i-U)-(vi-U) U is observed, hence $E[Y] = \gamma_0^{(CU)} + \gamma_A^{(CU)}A + \gamma_U^{(CU)}U + \gamma_{UA}^{(CU)}UA$.

When U is not observed, by standardisation we would hope that $E[\hat{\gamma}_A^{(j)}] \approx \delta_A = \gamma_A + \gamma_{UA}\pi_U$ (for $j = 0,1,2,C$), which represents the unconditional causal effect of A on Y. In MI(R*A), the $(1 - R_A)$ term may act as a partial proxy for U, therefore inclusion of the interaction means we expect $\widehat{E[\gamma_A^{(3)}]}$ to lie between the unconditional and marginal causal effects of A on Y.

For the cases where U is observed, we hope that $E[\hat{\gamma}_A^{(j)}] \approx \gamma_A$ for all models ($j = 0U, 1U, 2U, 3U, CU$), since the interaction with U is always modelled.

Where present in our models, we hypothesise that the $\hat{\gamma}_R^{(j)}$ and $\hat{\gamma}_{RA}^{(j)}$ terms may indicate MNAR when they are nonzero.

Analytical comments

It is instructive to consider a special case of scenario (ii) (see Figure 1), in which $R_A = 1 - U$. Suppose further that the true underlying regression model has no interaction, i.e. $E[Y] =$

$\gamma_0 + \gamma_A A + \gamma_U U$. Performing multiple imputation for A and including a missing indicator $1 - R_A$ in the outcome model – which corresponds to the MI(R+A) approach described above (model 2) - would be expected to perform well, as this analysis model matches the true model. Indeed, the regression coefficient of A on Y can be estimated without bias, $E[\hat{\gamma}_A^{(2)}] = E[\hat{\gamma}_A] = \gamma_A$. However, the model produces a biased estimate of the regression coefficient of Y on U, $E[\hat{\gamma}_R^{(2)}] = E[\hat{\gamma}_U] \approx \gamma_U \frac{\sigma_Y^2}{\gamma_A^2 \sigma_A^2 + \sigma_Y^2}$. This is because fitting the imputation model introduces regression dilution [16]. A justification is given in the Appendix. We emphasise that this should not be viewed as a shortcoming of multiple imputation, since multiple imputation in this case is targeting the regression coefficient of Y on A in the absence of missing data (completed data estimand), γ_A.

While the case $R_A = 1 - U$ may seem extreme, it could approximately hold in practice: for example, if a particular test (A) is commonly run if a particular unrecorded condition (U) is met, and is rarely run otherwise.

Simulation study set-up

The aims, general structure, and models, are described above. We consider the following specific data generating mechanisms, which cover all of the scenarios (i)-(vi) and (i-U)-(vi-U) described in Figure 1. We closely follow best practice for the design and reporting of simulation studies as proposed in [17].

For the $R_A \neq 1 - U$ case:

- We fix the sample size (number of observations within each simulation run) to be $n = 10{,}000$, and fix $\pi_U = 0.5$.
- We choose the intercepts as functions of the other parameters: α_0 such that $E[A] = 0$, γ_0 such that $E[Y] = 0$, and β_0 such that $P[R_A = 0]$ varies over the grid $\{0.25, 0.5, 0.75\}$.
- The main effect parameters, α_U, β_A, and γ_U are all varied over the grid $\{0, 0.1, 0.5, 1\}$, the parameter β_U over the grid $\{-1, 0, 0.1, 0.5, 1\}$ (a negative β_U is included to study whether the direction of correlation between U and R_A is important), while we fix $\gamma_A = 1$.
- The interaction effect parameters, β_{UA} and γ_{UA}, are varied between $\{0, 0.5\}$.
- The standard deviation of Y, σ_Y, is varied over the grid $\{0.1, 0.5, 1\}$, while we fix $\sigma_A = 1$.

For the $R_A = 1 - U$ case, we use the same simulation settings with the following exceptions:

- We exclude β_U, β_A and β_{UA}, which are redundant.
- We vary π_U over the grid $\{0.25, 0.5, 0.75\}$, as this is required to vary the proportion of missingness.

All combinations of the parameters are evaluated, resulting in 11808 scenarios, of which 288 cover the case where $R_A = 1 - U$.

For each scenario, we fit the models described in the previous section, and report estimates of the outcome coefficients from the various models. Each scenario is repeated 200 times and summary statistics over these iterations retained. For all parameters of interest – those

of the form $\hat{\gamma}_A^{(j)}$, $\hat{\gamma}_R^{(j)}$, and $\hat{\gamma}_{RA}^{(j)}$, we retain the 2.5th, 25th, 50th, 75th and 97.5th percentile parameter estimates. We also retain the average model-based standard errors and empirical standard errors for each parameter. For the parameters $\hat{\gamma}_A^{(j)}$ we also report the length and coverage of associated confidence intervals.

Results

Here we present a subset of the simulations that capture the main findings; full results are available – see *Availability of data and materials*. Throughout this section we restrict parameters to $\gamma_A = 1, \gamma_U = 1$, and $\sigma_Y = 1$, although we consider both $\gamma_{UA} = 0$ and $\gamma_{UA} = 0.5$. We also restrict to cases that result in $P[R_A = 1] = 0.5$. When $\gamma_{UA} = 0$, the marginal causal effect of A on Y, and the conditional causal effect of A on Y (when $U = 0$ and when $U = 1$) are 1. In this case, complete case analysis agrees closely with the completed data estimates. Most of our results focus on this case, for simplicity. When $\gamma_{UA} = 0.5$, by standardisation the marginal causal effect of A on Y, throughout, is 1.25, while the conditional causal effects of A on Y given $U = 0$ and $U = 1$ are 1 and 1.5 respectively. Qualitatively similar results were found when varying the remaining parameters in the outcome model and proportion of missing data. Results are summarized in figures and tables. The tables include mean estimates, average confidence interval width, and coverage for a targeted value of 1 in all cases.

Figure 2 and Table S1 show results for Scenarios (i)-(iii), i.e. where $\beta_A = \beta_{UA} = 0$. In addition, for this figure, we fix $P[R_A = 1] = 0.5, \gamma_A = 1, \gamma_U = 1, \gamma_{UA} = 0$ and $\sigma_A = 1$. Scenario (i) is the case where $\beta_U = \alpha_U = 0$. For Scenario (ii), β_U controls the strength of the

relationship between U and R_A, with the extreme case $R_A = 1 - U$, with $\alpha_U = 0$. For Scenario (iii), α_U additionally controls the strength of the relationship between U and A – i.e. introduces unmeasured confounding.

The causal effect of A on Y is 1 (dotted vertical line in leftmost panels). For $\beta_U = \alpha_U = 0$ all methods' estimates of γ_A are able to recover this without bias and with appropriate coverage. As β_U increases, all methods are still able to estimate the causal effect well, except that MI(A) becomes biased when $R_A = 1 - U$. As α_U increases, the completed data model becomes biased because of unmeasured confounding. We see that the MIMI approaches and complete case analysis are able to mitigate this to some extent, and successfully when $R_A = 1 - U$. The estimate of γ_R becomes nonzero for the MIMI methods when $\alpha_U \neq 0$: it is through this that the MIMI methods are able to partly correct for the unmeasured confounding. Note that when $\beta_U = -1$ then the $\hat{\gamma}_R^{(j)}$s are negative rather than positive, but the $\hat{\gamma}_A^{(j)}$s are similar to the $\beta_U = 1$ case.

Figure 3 and Table S2 present the same scenarios as Figure 2 but with $\gamma_{UA} = 0.5$; hence the marginal and conditional causal effects of A on Y differ, as explained above. In the $R_A = 1 - U$ case, with $\alpha_U = 0$ the completed data model estimates $\delta_A = \gamma_A + \pi_U \gamma_{UA} = 1.25$, the marginal effect, while complete case analysis estimates the conditional effect when $U = 0$ (which is $\gamma_A = 1$); this is of course not surprising as there is only data when $U = 0$. The MI(R*A) estimate agrees with the complete case, while MI(A) and MI(R+A) tend to interpolate between the two. However, when $\beta_U = -1$ we now see that all methods have increased bias. This is because the reversal of the correlation between U and R_A means that missingness is more likely when $U = 1$, when the conditional causal effect is 1.5.

Figure 4 and Table S3 show results for scenario (i-U)-(iii-U) with $\gamma_{UA} = 0$, i.e. the same conditions as Figure 2 except that U is now measured. In this MAR case, the missing indicator should be redundant, and the concern is that its inclusion may introduce bias into the estimates. We see that all methods perform well in recovering γ_A. MI(A) and MI(R+A) are in almost perfect agreement.

Figure 5 and Table S4 show results for scenarios (iv) and (v), where we examine the effects of missingness in A depending on A itself. Here, the scenario dictates that $\alpha_U = \beta_{UA} = 0$, and we additionally fix $\gamma_A = 1, \gamma_U = 1, \gamma_{UA} = 0, \sigma_Y = 1$ and $\sigma_A = 1$. The key varying parameters are β_A, which controls the dependence of R_A on A, and β_U, which controls the dependence of R_A on U.

When $\beta_U = 0$ (corresponding to scenario (iv)), MI(A) is biased in estimating γ_A. However, MI(R+A) and MI(R*A) are not biased. When $\beta_U \neq 0$ things are more complicated, and there is no clear approach that minimizes the bias. What is consistent, however, is that γ_R estimates are nonzero when either β_U or β_A are not zero.

Figure 6 and Table S5 show results for Scenario (iv-U) and (v-U), which are the same scenarios as in Figure 5 except that U is measured. In these cases, changing values of β_U do

not cause particular problem for any method, while nonzero β_A introduces bias in estimation of γ_A for MI(A) only.

Figure 7 and Table S6 show results for Scenario (vi). This is the most flexible scenario with no constraints on the parameter values. Here we illustrate the case where $\gamma_A = 1, \gamma_U = 1, \gamma_{UA} = 0.5, \sigma_A = 1, \sigma_Y = 1$ and $\beta_{UA} = 0$, and $\alpha_U = 0.5$.

The results are similar to those for Scenario (v) except that γ_A is more commonly overestimated.

Further results are given in the Supplements: Figures S1-S3 and corresponding tables S7-S9.

Discussion

In this paper we have explored, through simulation, the potential merits of supplementing multiple imputation with a missing indicator, particularly in circumstances where missingness is not at random, and the missingness may moreover act as a proxy for unmeasured confounding. We emphasise that, in contrast to the usual missing data literature that targets completed data estimands, here we target causal estimands that are not available in general even with completed data (because of unmeasured confounding). In scenarios where missingness of an exposure depends on an unmeasured confounder, the missingness indicator can be used as a proxy for the unmeasured confounding, and this may reduce bias in some situations. Careful consideration of the likely missingness

mechanisms for a given clinical question/ dataset is key to deciding on the analytical approach.

In the MCAR and MAR scenarios, without unmeasured confounding, adding a missing indicator to multiple imputation did not introduce bias in estimation of causal effects. In the MNAR scenarios without unmeasured confounding, adding a missing indicator generally reduced bias compared with multiple imputation alone. In the presence of unmeasured confounding, bias in estimation was sometimes better and sometimes worse when including a missing indicator and/or its interaction with the main effect, depending on the relationships between the parameters. This reflects the potentially complex relationships, and shows that care should be given, and decisions based on a study-by-study basis. In all cases, when unmeasured confounding and/or MNAR exists, the missing indicator coefficient and/or its interaction with the main effect coefficient were estimated to be non-zero. These non-zero effect estimates of the missing indicators act as a signal that there may be MNAR mechanisms present, and hence it would be difficult or impossible to obtain unbiased causal effects. Any disagreement between the main effect parameter estimates with and without including a missing indicator provide a similar indication.

The 'missing indicator' approach has a somewhat negative reputation in the causal inference literature. This is because it is usually coupled with a weak approach to impute the missing data itself - such as using the unconditional mean [8]. With such application, missing indicator is known to lead to biased estimation even under MCAR [9–12]. The idea of combining the missing indicator approach with multiple imputation was first proposed by [6], and has been further explored by [7] and [4]. In those articles, the focus is on

handling missing data in covariates used in propensity scores, whereas here we consider missing data in the exposure of interest. Nevertheless, [4] in particular noted that the use of missing indicators can partly adjust for unmeasured confounding, similar to our findings. However, we find that they can also make the situation worse, so considerable care is needed, on a study-by-study basis.

Strengths and limitations

We explored a wide range of simulation settings in a fully factorial design. While we can only present a limited range of results in the paper, the simulation code and results are available online for inspection. Nevertheless, simulations are necessarily simpler than scenarios that might be encountered in practice. First, missingness may affect many covariates. While addition of missing indicators, and interactions, seems robust, it may break down in some scenarios with complex multivariate patterns of missingness, and may also lead to unacceptable model complexity. Second, there may be multiple unmeasured or partially measured confounders. However, we could consider multiple confounders as being summarized by a propensity score, for example, and thus we expect the results here to generalize to the multiple confounders case. We emphasise that we focused on missing data in exposure where the causal estimand rather than the completed data estimand was targeted, and that results here should not be generalized to different scenarios [18]. Finally while we have presented limited analytical findings in this paper, it is likely that the bias formulas could be derived for a wide range of the scenarios presented, which we leave as a topic for further research.

Conclusions

We recommend that addition of a missing indicator, and corresponding interaction terms, can supplement, but not replace, standard multiple imputation. In particular, we recommend the use of MIMI (including interactions between missing indicators and the corresponding variable) as a strategy for handling missing data in causal effect estimation problems. Non-zero estimates of the missing indicator then alert to possible occurrence of MNAR and/or unmeasured confounding, and the need for further sensitivity analysis. We caveat that the use of missing indicators should not replace careful consideration of assumed plausible causal structures, and drawing a causal diagram to depict these assumptions remains the starting point for a well-conducted causal inference.

Abbreviations

MAR: Missing at random

MCAR: Missing completely at random

MI: Multiple imputation

MIMI: Multiple imputation with missing indicator

MNAR: Missing not at random

Declarations

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All simulation results are available at *https://figshare.com/articles/Output_from_simulations/10320617*, and the code is available at *https://github.com/mattsperrin/missing_indicator_sim_paper*.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partially funded by the MRC-NIHR Methodology Research Programme [grant number: MR/T025085/1]. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

MS: design of study, perform simulation study, draft paper. GM: design of study, major contributions to editing paper. Both authors read and approved the final manuscript.

Acknowledgements

The authors thank Thomas House for useful discussions.

Appendix: Bias for estimating γ_u from imputed data when $R_A = 1 - U$

Here we give an informal justification for the bias result. In this section we use the superscript $*$ to denote true values of parameters.

First consider the imputation model

$$a_i = \phi_0^* + \phi_Y^* y_i + \delta_i,$$

for $\mathcal{J} = \{i: R_{A,i} = 1\}$, i.e. non-missing As, with $\delta_i \sim N(0, \tau^2)$. Note that u_i does not appear in this model because $u_i = 0$ for all $i \in \mathcal{J}$.

Now $y_i \sim N(\mu_{Y,i} = \gamma_0^* + \gamma_A^* a_i, \sigma_Y^2)$ for $i \in \mathcal{J}$. In analogy with p175 of [16], consider a hypothetical imputation model based on $\mu_{Y,i}$ instead of y_i:

$$a_i = \psi_0^* + \psi_Y^* \mu_{Y,i} + \xi_i,$$

from which it is apparent that $\psi_Y^* = 1/\gamma_A^*$. Additionally, across all observations, $\mathrm{Var}(\mu_Y) = \gamma_A^{*2}\sigma_A^{*2}$

In analogy with [16](referencing [19]) we have that

$$\phi_Y^* = \frac{\psi_Y^* \gamma_A^{*2} \sigma_A^{*2}}{\gamma_A^{*2}\sigma_A^{*2} + \sigma_Y^{*2}} = \frac{\gamma_A^* \sigma_A^{*2}}{\gamma_A^{*2}\sigma_A^{*2} + \sigma_Y^{*2}}.$$

Moreover, $\phi_0^* = 0$ because A and Y are both centred.

The imputation model is then used to impute values for the missing a_is; i.e. for $i \notin \mathcal{J}$, if we knew the true imputation model,

$$a_{i,imp} = \phi_0^* + \phi_Y^* y_i + \delta_i = \frac{\gamma_A^* \sigma_A^{*2}}{\gamma_A^{*2}\sigma_A^{*2} + \sigma_Y^{*2}} (\gamma_A^* a_i + \gamma_U^* u_i + \epsilon_i) + \delta_i.$$

Now consider again the outcome model,

$$y_i = \gamma_0^* + \gamma_A^* a_i + \gamma_U^* u_i + \epsilon_i.$$

In the absence of missing data, we would of course simply solve using least squares, and if $\gamma = (\gamma_0, \gamma_A, \gamma_U)$ and $\hat{y}_i(\gamma) = \gamma_0 + \gamma_A a_i + \gamma_U u_i$, then $\tilde{\gamma} = \text{argmin}(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)$, then of course $E_Y[\tilde{\gamma}_U] = \gamma_U^*$.

As we have missing data, rewriting the outcome model to replace the missing a_is with their imputed versions, for substitution into the least squares formula we have:

$$\hat{y}_i(\gamma) = \gamma_0 + \gamma_A((1 - u_i)a_i + u_i a_{i,imp}) + \gamma_U u_i.$$

The residual sum of squares can then be written as

$$\hat{\gamma} = \text{argmin} \sum_{i=1}^{n} \{(\gamma_0^* - \gamma_0) + (\gamma_A^* - \gamma_A)a_i + (\gamma_U^* - \gamma_A\gamma_U^*\kappa - \gamma_U)u_i + (-\gamma_A\kappa\epsilon_i - \gamma_A\delta_i)u_i + (\gamma_A - \gamma_A\kappa\gamma_A^*)u_i a_i\}^2,$$

where $\kappa = \frac{\gamma_A^*\sigma_A^{*2}}{\gamma_A^{*2}\sigma_A^{*2} + \sigma_Y^{*2}}$ is a constant.

To consider minimising this expression, consider each bracket in turn. To minimise the first bracket, it is clear that $E_Y[\hat{\gamma}_0] = \gamma_0^*$. It is also apparent that $E_Y[\hat{\gamma}_A] = \gamma_A^*$, since we must minimise the second bracket, and the fourth and fifth brackets are additional error contributed by the imputed data, which cannot be reduced. This leaves the third bracket, which is minimised at

$$E_Y[\gamma_U^* - \hat{\gamma}_A\gamma_U^*\kappa - \hat{\gamma}_U] = 0.$$

Rearranging yields,

$$E_Y[\hat{\gamma}_U] = \gamma_U^* \frac{\sigma_Y^{*2}}{\gamma_A^{*2}\sigma_A^{*2} + \sigma_Y^{*2}},$$

as claimed.

References

1. Rubin DB. Inference and missing data. Biometrika. 1976;63:581–92.

2. Mohan K, Pearl J, Tian J. Graphical Models for Inference with Missing Data. Adv Neural Inf Process Syst. 2013;26 December:1–9.

3. Hernan MA, Robins JM. Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming; 2019.

4. Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. Eur J Epidemiol. 2019;34:23–36.

5. Little RJA. Regression with missing X's: A review. J Am Stat Assoc. 1992.

6. Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. Stat Med. 2009;28:1402–14.

7. Seaman S, White I. Inverse Probability Weighting with Missing Predictors of Treatment Assignment or Missingness. Commun Stat Methods. 2014;43:3499–515.

8. Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel.

Biostatistics. 2018.

9. Jones MP. Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. J Am Stat Assoc. 1996;91:222–30.

10. Knol MJ, Janssen KJM, Donders ART, Egberts ACG, Heerdink ER, Grobbee DE, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. J Clin Epidemiol. 2010;63:728–36.

11. Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. CMAJ. 2012;184:1265–9.

12. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol. 1995;142:1255–64.

13. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. Stat Methods Med Res. 2012;21:243–56.

14. Tilling K, Williamson EJ, Spratt M, Sterne JAC, Carpenter JR. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. J Clin Epidemiol. 2016;80:107.

15. Rubin D. Multiple imputation for nonresponse in surveys. 2004.

16. Frost C, Thompson SG. Correcting for regression dilution bias: Comparison of methods for a single predictor variable. J R Stat Soc Ser A Stat Soc. 2000;163:173–89.

17. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38:2074–102.

18. Little R. On algorithmic and modeling approaches to imputation in large data sets. Stat Sin. 2019.

19. Snedecor GW, Cochran WG. Statistical Methods. 6th Edition. Appl Stat. 1968.

Figure Legends

Figure 1: Causal directed acyclic graphs denoting missingness mechanism for A, R_A: six scenarios considered in the paper.

Figure 2: Results for scenarios (i)-(iii), with $\gamma_{UA} = 0$. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U, with the special case $R_A = 1 - U$ on the top row. Within each graph, the y-axis varies α_U.

Figure 3: Results for scenarios (i)-(iii), with $\gamma_{UA} = 0.5$. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U, with the special case $R_A = 1 - U$ on the top row. Within each graph, the y-axis varies α_U.

Figure 4: Results for scenarios (i-U)-(iii-U), with $\gamma_{UA} = 0$. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U, with the special case $R_A = 1 - U$ on the top row. Within each graph, the y-axis varies α_U.

Figure 5: Results for scenarios (iv) and (v), with $\gamma_{UA} = 0$. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies β_A.

Figure 6: Results for scenarios (iv-U) and (v-U), with $\gamma_{UA} = 0$. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies β_A.

Figure 7: Results for scenario (vi), with $\gamma_{UA} = 0$ and $\alpha_U = 0.5$. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies β_A.

Supplementary File List

Supplemental Material

Figures

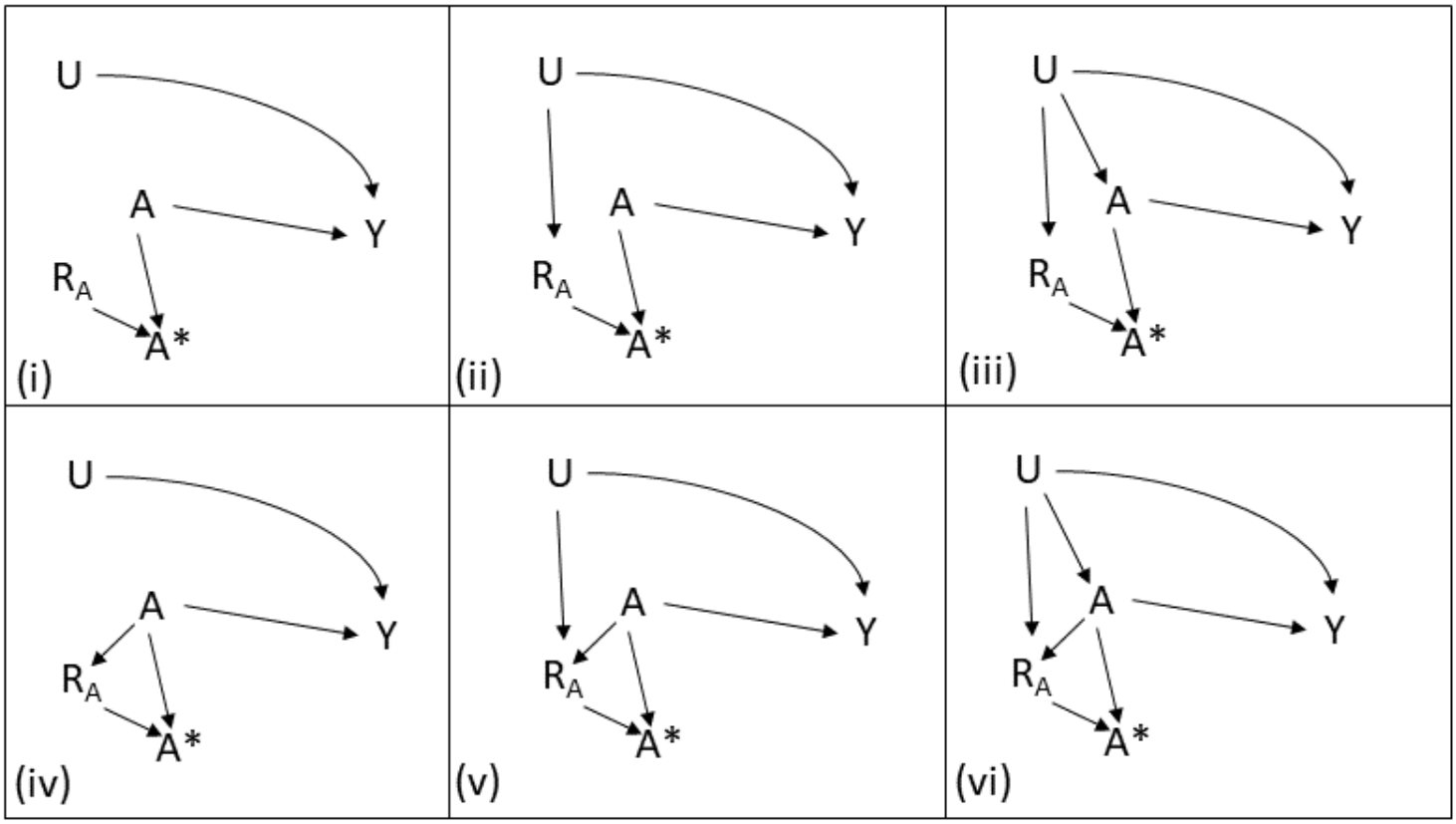

Figure 1

Causal directed acyclic graphs denoting missingness mechanism for A, R_A: six scenarios considered in the paper.

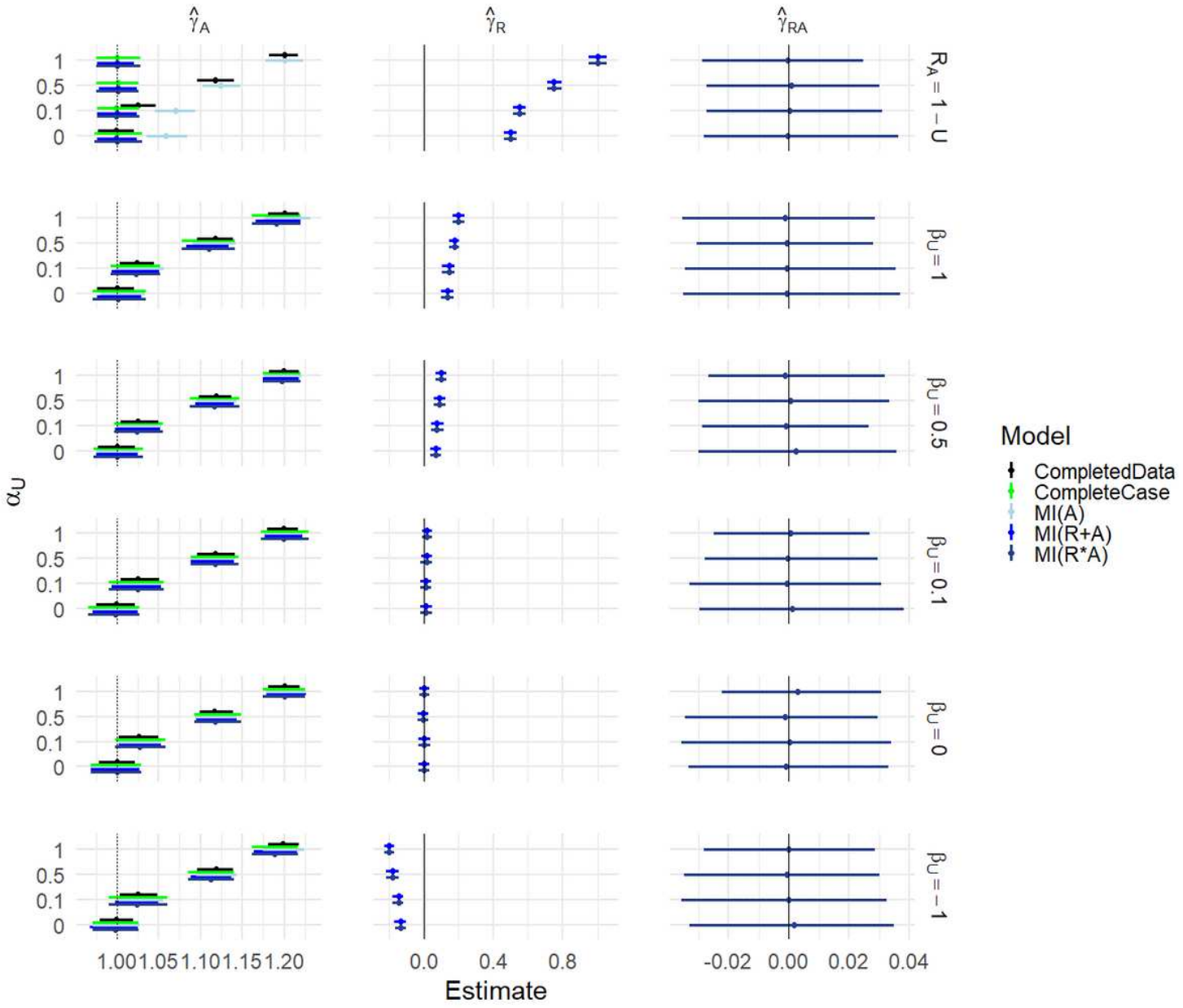

Figure 2

Results for scenarios (i)-(iii), with γ_UA=0. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U, with the special case R_A=1-U on the top row. Within each graph, the y-axis varies α_U.

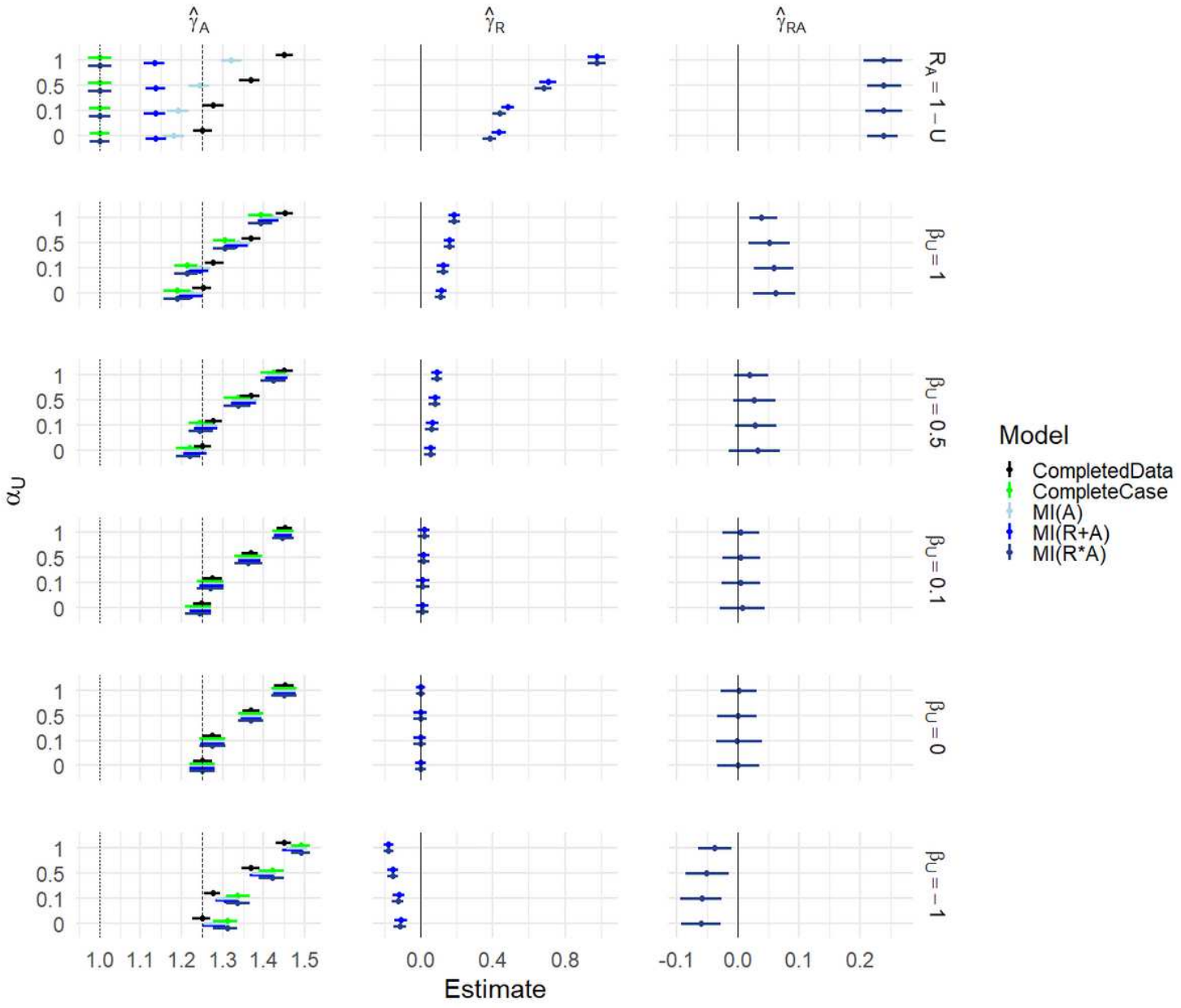

Figure 3

Results for scenarios (i)-(iii), with γ_UA=0.5. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U, with the special case R_A=1-U on the top row. Within each graph, the y-axis varies α_U.

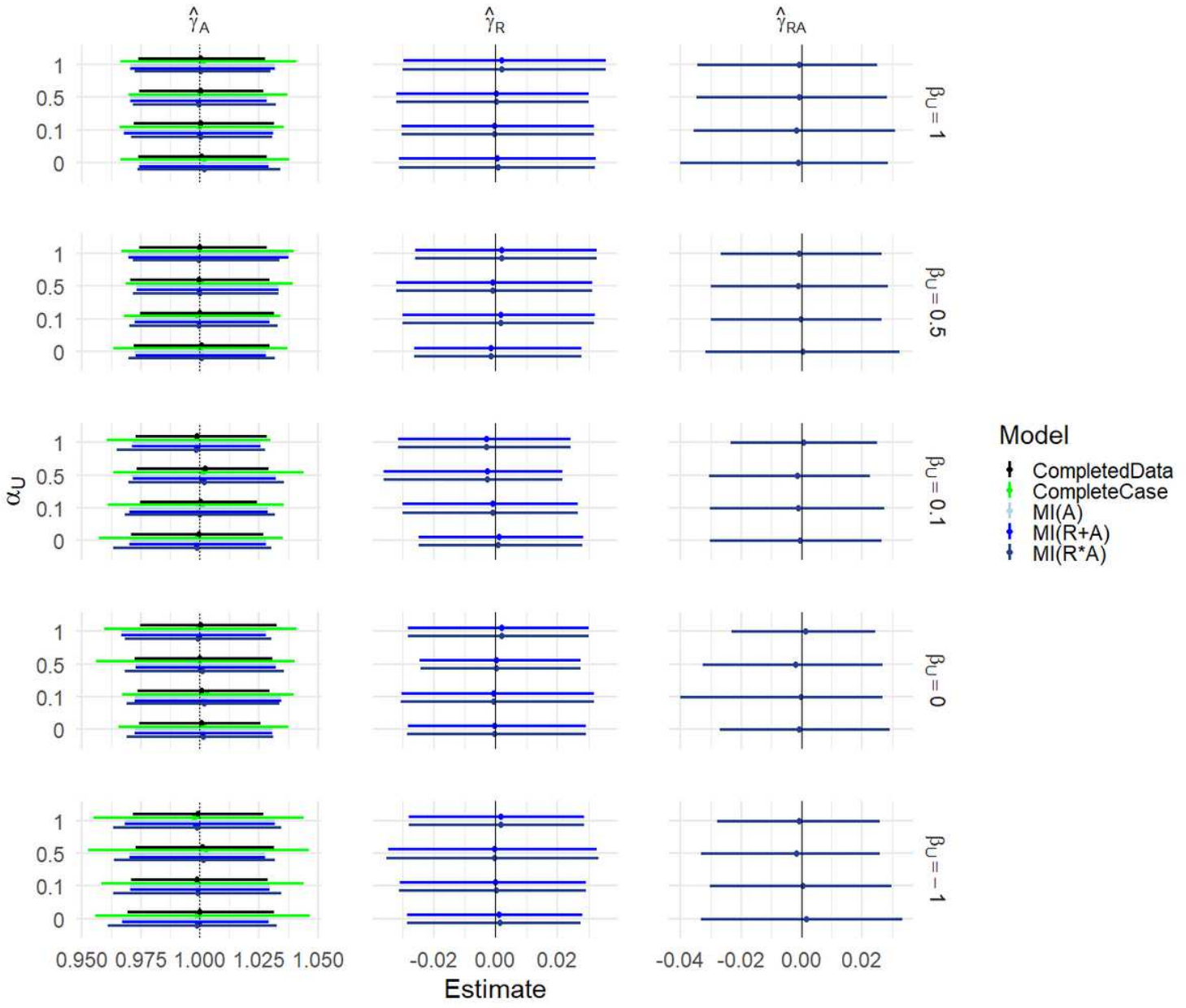

Figure 4

Results for scenarios (i-U)-(iii-U), with γ_UA=0. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U, with the special case R_A=1-U on the top row. Within each graph, the y-axis varies α_U.

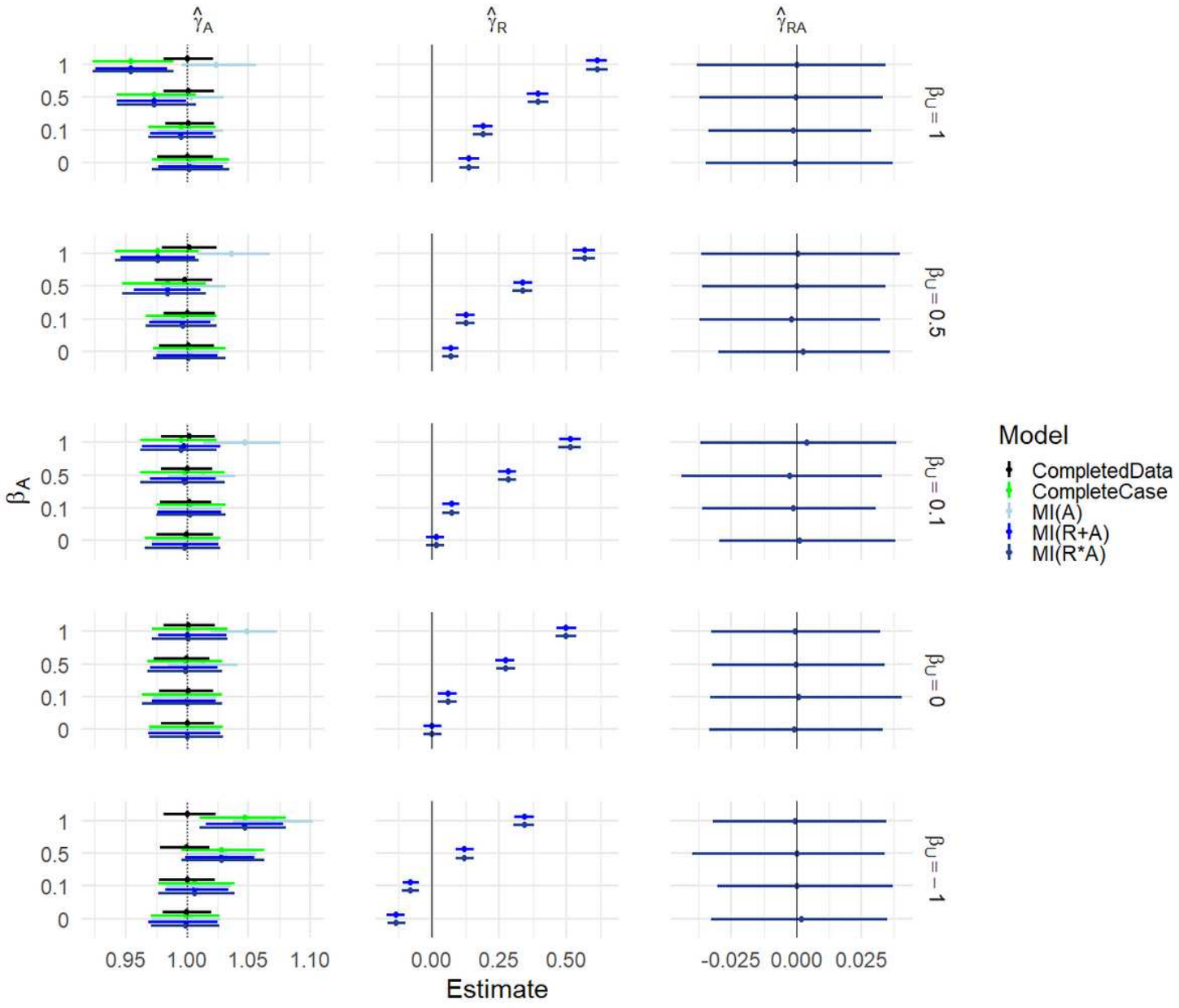

Figure 5

Results for scenarios (iv) and (v), with γ_UA=0. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies β_A.

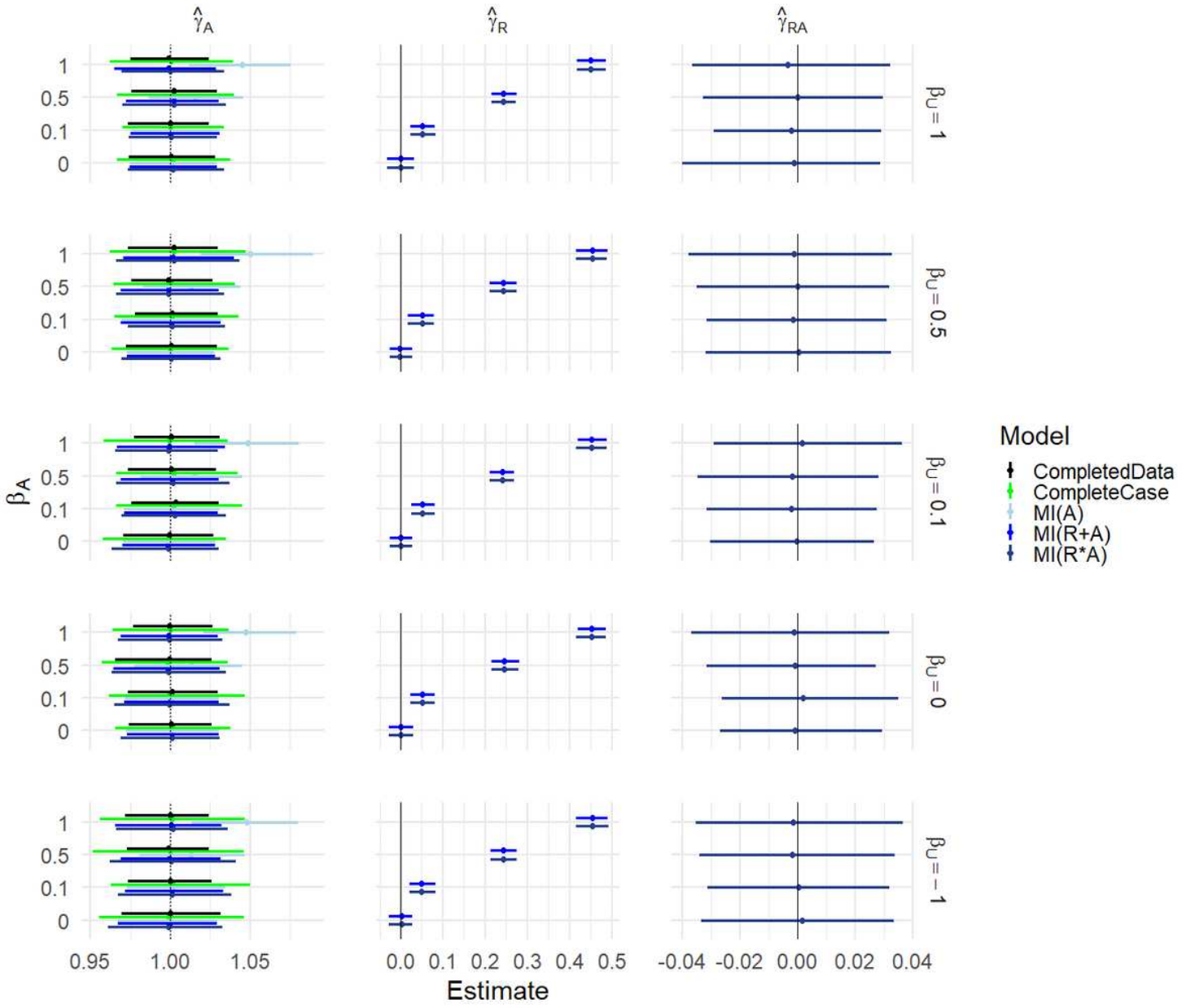

Figure 6

Results for scenarios (iv-U) and (v-U), with γ_UA=0. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies β_A.

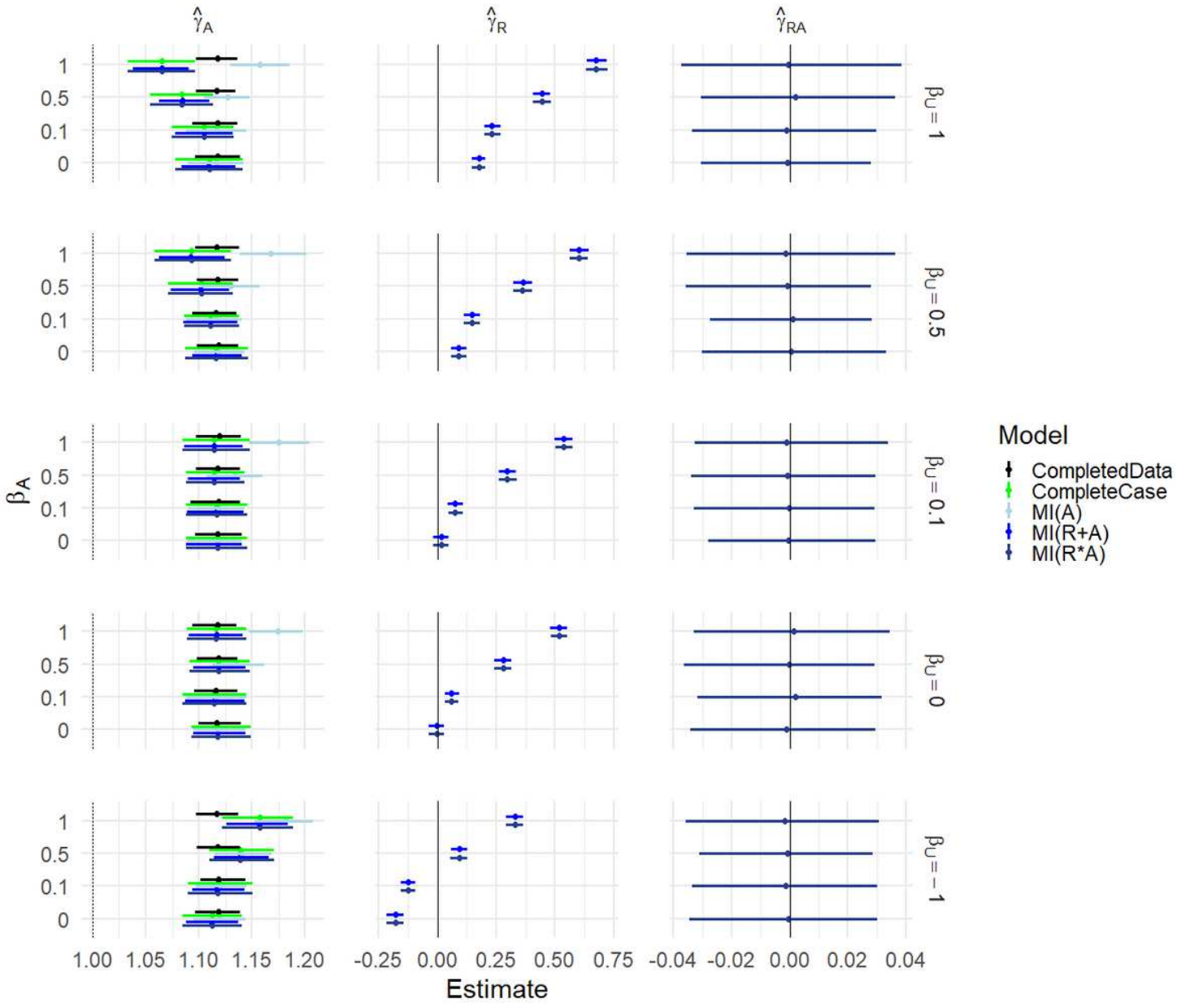

Figure 7

Results for scenario (vi), with γ_UA=0 and α_U=0.5. Mean of estimated coefficients across simulations; error bars represent the 2.5th and 97.5th percentiles. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies β_A.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- miucpapersuppmaterialR2.docx