Use of 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 when used inappropriately, its use in conjunction with other imputation approaches may unlock the potential value of missingness to reduce bias and improve prediction.

Methods: We conducted a simulation study to determine when the use of a missing indicator, combined with an imputation approach, such as multiple imputation, would lead to improved model performance, in terms of minimising bias for causal effect estimation, and improving predictive accuracy, under a range of scenarios with unmeasured variables. We use directed acyclic graphs and structural models to elucidate causal structures of interest. We consider a variety of missingness mechanisms, then handle these using complete case analysis, unconditional mean imputation, regression imputation and multiple imputation. In each case we evaluate supplementing these approaches with missing indicator terms.

Results: For estimating causal effects, we find that multiple imputation combined with a missing indicator gives minimal bias in most scenarios. For prediction, we find that regression imputation combined with a missing indicator minimises mean squared error.

Conclusion: In the presence of missing data, careful use of missing indicators, combined with appropriate imputation, can improve both causal estimation and prediction accuracy.

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

Background

Missing data is a common feature in observational studies. Particularly for studies that target causal effects, but also for prediction, careful thought is needed when deciding how to handle missing data. 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].

In the case of MCAR and MAR, an unbiased estimator of any causal effect of interest exists, with approaches such as complete case analysis providing an unbiased causal estimates under MCAR. One of the most popular means of handling missing data is multiple imputation, which can provide unbiased estimates under both MAR and MCAR. In contrast, under MNAR, unbiased estimators of a given causal effect may or may not exist. In this case, drawing causal diagrams or Directed Acyclic Graphs (DAGs) that include the missingness mechanism, called m-graphs [2], can help to identify whether a given effect can be estimated, and how to do so.

One form of MNAR is where the missingness mechanism depends on entirely unmeasured variables, that are causally related to the outcome variable. These unmeasured variables may act as confounders, in which case, unbiased estimators of a causal effect do not exist even in the absence of missing data.

An emerging hypothesis is that in some scenarios with such unmeasured confounding affecting a particular estimand, missing data may be a blessing rather than a curse. For example, in electronic health records, presence of a particular laboratory test result

indicates that a decision was made to run this test, and the reason for this decision is likely to depend on unrecorded health characteristics of the patient. These unrecorded characteristics may affect both the result of the laboratory test, and the outcome of interest. In these cases, the missingness mechanism (i.e. presence or absence of the laboratory test) may act as a proxy for unmeasured confounding, allowing for partial adjustment [3]. Although in MCAR and MAR scenarios, use of missing indicators can introduce bias [4,5], use of missing indicators may reduce bias where the missingness is informative, and particularly when the missing indicator is used in conjunction with multiple imputation (MIMI) [3,6–8].

We hypothesise that use of missing indicator in conjunction with regression/ multiple imputation methods in such cases would also improve predictive accuracy. Indeed, we suggest that the approach to imputation that one utilises should differ depending on the underlying analytical aim. Specifically, how one handles missing data should arguably differ for studies aiming to estimate causal effects (where primary interest is in obtaining unbiased parameter estimates) as opposed to studies aiming to develop risk models for a particular prediction task (where primary interest is in obtaining accurate risk predictions, regardless of the underlying parameter estimates).

In this paper we study, through simulation supplemented with analytical findings, the potential for using the missingness mechanism to partly adjust for unmeasured confounding, and study the scenarios where this can be beneficial, both for causal and prediction objectives.

Methods

Scenarios and data generating mechanisms

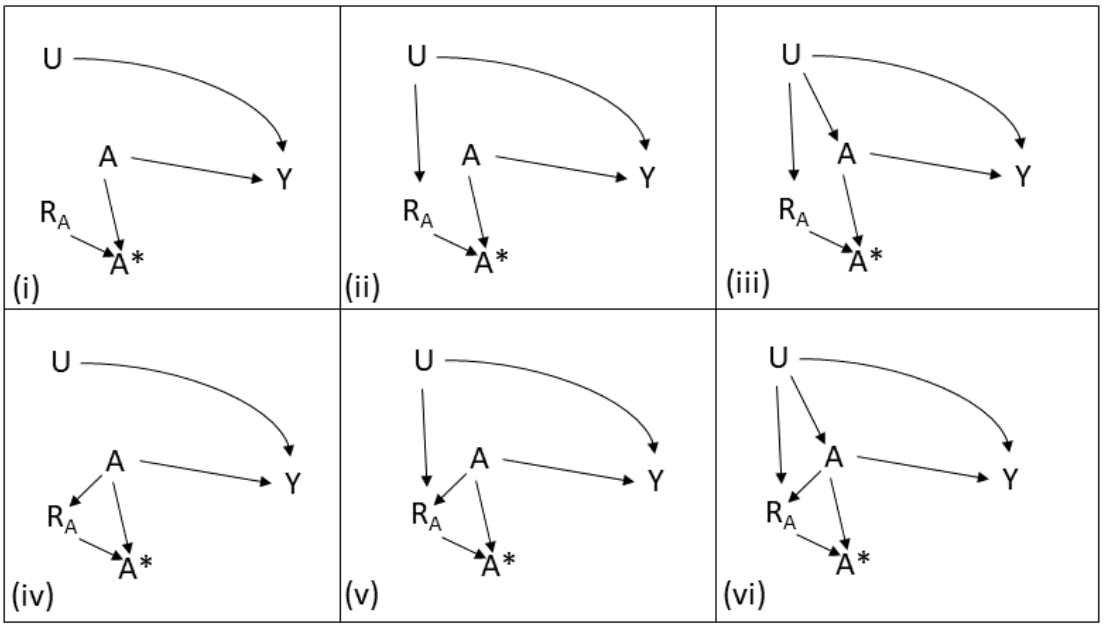

Figure 1: Causal DAGs denoting missingness mechanism (i.e. m-graphs) for A, R_A: six scenarios considered in the paper.

Our primary aim is to identify missing data strategies that recover causal effects with minimal bias in a variety of MNAR scenarios; as a secondary aim we will examine predictive accuracy of the models, representing the case where ones primary interest is developing a prediction model, rather than recovering causal effects. The scenarios that we will consider in this paper are given in Figure 1. We consider a partially observed exposure A, a fully observed outcome Y and an unobserved variable U. The missingness mechanism for A is R_A where $R_A = 1$ when A is missing (in the example in the Introduction, R_A might denote a laboratory test being performed or not). A^* is the observable part of A, i.e. $A^* = A$

when $R_A = 0$, and missing when $R_A = 1$. So A^* is what we observe, while A includes unobserved values.

Scenario (i) corresponds to MCAR, since R_A does not have any inward arrows. All other scenarios, (ii)-(vi), are MNAR as R_A is causally affected by U or A or both. We do not consider MAR scenarios, which are well studied elsewhere in the literature. In scenarios (i) and (iv) complete case analysis is unbiased (see e.g. [9] 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 would 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 unmeasured U. It therefore seems reasonable to include R_A in the outcome model. This may reduce bias in the estimation of the effect of A on Y, and provide at least partial information about the effect of U on Y.

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

For consideration of causal effects we will use the counterfactual notation, 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 [10] for an introduction to causal inference with counterfactuals. Our primary aim is to recover the causal effect of A on Y, i.e. $E[Y(A = 1) - Y(A = 0)]$, and we also have a secondary interest in: 1) the causal effect of U on Y, $E[Y(U = 1) - Y(U = 0)]$, and 2) the mean squared error (MSE) of the resulting model, MSE $= n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, where there are n observations indexed by i, and $\hat{y}_i$ is the predicted value of the outcome y_i.

Considered approaches

We consider the following imputation and modelling approaches.

First, a complete case analysis. This fits the model $E[Y] = \hat{\gamma}_0^0 + \hat{\gamma}_A^0 A^*$, restricting to observations where $R_A = 0$.

Second we consider (unconditional) mean imputation, where missing As are simply replaced by the unconditional mean of the observed As.

Finally, we consider regression imputation and multiple imputation, each of which require an imputation model of the form $E[A^*] = \phi_0 + \phi_Y Y$. For regression imputation, missing As are replaced by their predicted mean from this model, $a_i = \hat{\phi}_0 + \hat{\phi}_Y y_i$. For multiple imputation, missing As are imputed by a random draw from the predictive distribution implied by the imputation model, $a_i \sim N(\hat{\phi}_0 + \hat{\phi}_Y y_i, \tau_i^2)$, where τ_i^2 is the sum of the variances from the error in the mean and the residual variance. 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}. For each imputation approach, we consider the following three outcome/analysis models:

1. $E[Y] = \hat{\gamma}_0^1 + \hat{\gamma}_A^1 A_{\text{imp}}$.

2. $E[Y] = \hat{\gamma}_0^2 + \hat{\gamma}_A^2 A_{\text{imp}} + \hat{\gamma}_R^2 R_A$.

3. $E[Y] = \hat{\gamma}_0^3 + \hat{\gamma}_A^3 A_{\text{imp}} + \hat{\gamma}_R^3 R_A + \hat{\gamma}_{RA}^3 A_{\text{imp}} R_A$.

When A_{imp} are generated using multiple imputations, Model 1 represents a standard multiple imputation approach, while models 2 and 3 are variants of the MIMI approach. In model 3, if we view R_A as a proxy for U we would hope that $\hat{\gamma}_A^3 \approx \gamma_A$, $\hat{\gamma}_R^3 \approx \gamma_U$, and $\hat{\gamma}_{RA}^3 \approx \gamma_{UA}$. For the other models, by standardisation, we would hope that $\hat{\gamma}_A^j \approx \gamma_A + \gamma_{UA}\pi_U$ (for $j = 0,1,2$), which represents the marginal causal effect of A on Y. For each model, an estimate of each parameter can be produced under each of the three imputation approaches we consider, with the exception that Model 3 is not identified for unconditional mean imputation.

Analytical comments

It is instructive to consider the special case of scenario (ii) (see Figure 1) where $R_A = U$. While this may seem extreme, it could well happen in practice: for example, if a particular blood test is only run if a particular condition is met, and that condition is not recorded. In

this case, it is trivial that the causal effect of U on Y is recoverable. In both scenarios we have exchangeablity $U \perp\!\!\!\perp Y(u)$, therefore

$$P[Y(u)] = P[Y(u)|U = u] = P[Y|U = u] = P[Y|R_A = u],$$

where the equalities follow, respectively, from $U \perp\!\!\!\perp Y(u)$, then consistency, and finally that $R_A = U$. This also holds in Scenario (iii).

Interestingly, if we impute the exposure A through multiple imputation, then fit an outcome model with the imputed A, A_{imp}, and include R_A, then when $U = R_A$, this model produces a biased estimate of the effect of U on Y even in the simple case of Scenario (ii) with $\gamma_{UA} = 0$, so that U and A do not interact in the outcome model.

In such a case the estimate produced has $E[\hat{\gamma}_U] \approx \gamma_U \frac{\sigma_Y^2}{\gamma_A^2\sigma_A^2+\sigma_Y^2}$. This is because fitting the imputation model introduces regression dilution in this case [11]. A justification is given in the Appendix.

Simulation study

Simulation 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) described in Figure 1. We closely follow best practice for the design and reporting of simulation studies as proposed in [12].

For the $R_A \neq U$ 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 $P[R_A = 1]$ varies over the grid $\{0.25, 0.5, 0.75\}$.
- The main effect parameters, $\alpha_U, \beta_U, \beta_A$, and γ_U are all varied over the grid $\{0, 0.1, 0.5, 1\}$, 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 $R_A = U$ 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 9504 scenarios, of which 288 cover the case $R_A = U$.

For each Scenario we fit the models described in the previous section, and report estimates of the outcome coefficients from the various models, using each method of imputation. Each scenario is repeated 200 times and summary statistics over these iterations retained. For the parameters $\hat{\gamma}_A^0, \hat{\gamma}_A^1, \hat{\gamma}_A^2, \hat{\gamma}_R^2, \hat{\gamma}_A^3, \hat{\gamma}_R^3$, and $\hat{\gamma}_{RA}^3$ we retain the 2.5th, 25th, 50th, 75th and 97.5th percentile parameter estimates. We retain the same percentiles for the mean

squared error of each model fit. We also retain the average model-based standard errors and empirical standard errors for each parameter.

Simulation results

Here we present a subset of the simulations that capture the main findings. For ease of interepretation, throughout this section we restrict to the outcome model with $\gamma_A = 1$, $\gamma_U = 1, \gamma_{UA} = 0.5$, although σ_Y is varied. We also restrict to cases that result in $P[R_A = 1] = 0.5$. As a result, 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. This restriction is made for ease of interpretation. Note that if $\gamma_{UA} = 0$, we observe the unfaithfulness property: the marginal causal effect, and two conditional causal effects agree; as a result, complete case analysis is always unbiased. Otherwise, qualitatively similar results were found when varying the γ parameters in the outcome model and proportion of missing data.

Figure 2 shows results for Scenarios (i) and (ii), i.e. where $\alpha_U = \beta_A = \beta_{UA} = 0$. In addition we fix $P[R_A = 1] = 0.5, \gamma_A = 1, \gamma_U = 1, \gamma_{UA} = 0.5$ and $\sigma_A = 1$. For scenario (ii), β_U controls the strength of the relationship between U and R_A, with $\beta_U = 0$ for Scenario (i).

In these scenarios the marginal causal effect of A on Y is 1.25 (dashed vertical line in leftmost panels); we expect the γ_A estimates to target this for all models except model 3, as this explicitly models the interaction, so should target the conditional causal effect when $U = 0$, which is 1 (dotted vertical line in leftmost panels). In fact, we see for Scenario (i) that the marginal causal effect is estimated well by all models, except regression

imputation with model 1 or model 2 that have higher bias of the marginal effect. In Scenario (ii) we see the bias increasing for all approaches as β_U increases. As a proxy for γ_U, estimates $\hat{\gamma}_R$ are a substantial underestimate. For estimating the interaction term, model 3 following multiple imputation estimates a value close to zero while model 3 following regression imputation estimates a larger value. Finally, considering the MSE, we see that regression imputation has substantially lower MSE regardless of the fitted outcome model, but lowest in Model 3, while mean imputation has a higher MSE than the other approaches.

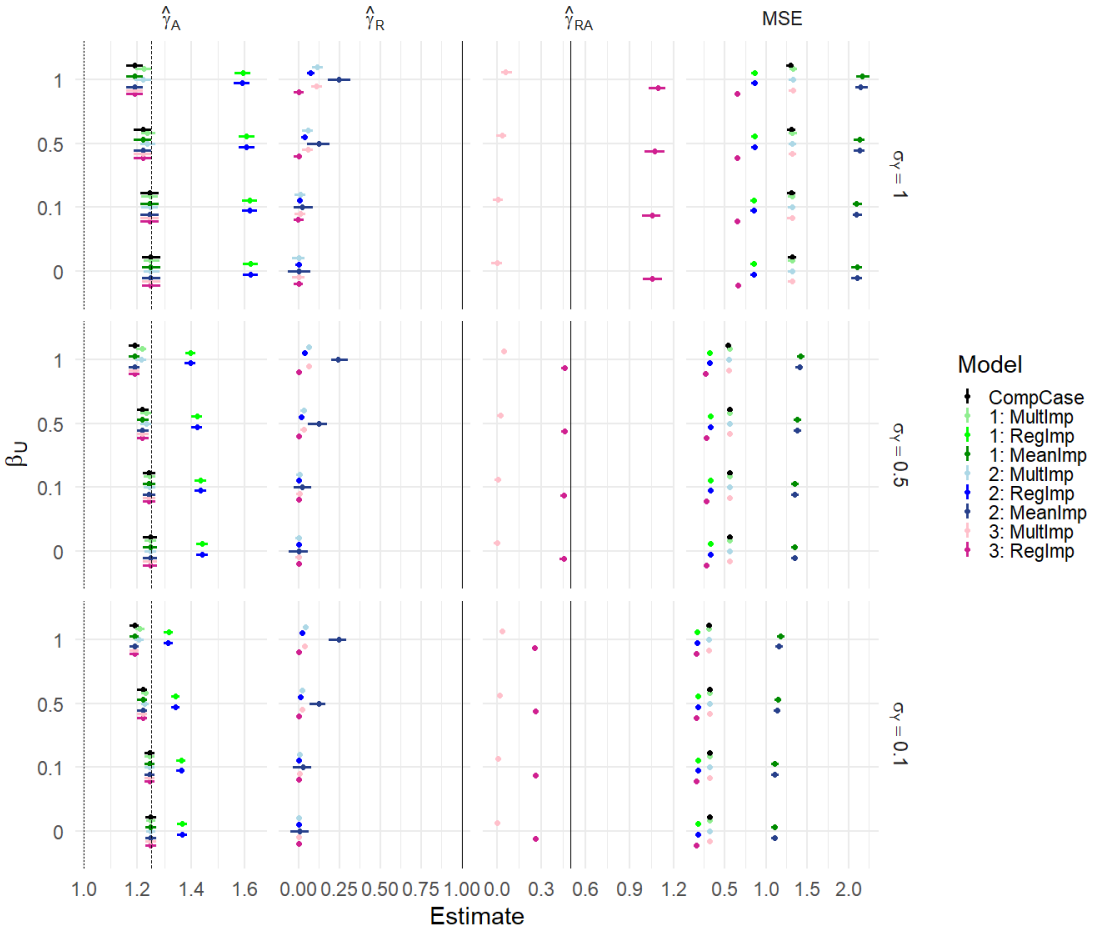

Figure 2: Results for scenarios (i) and (ii). Estimates with 95% empirical confidence intervals. Columns are different parameter estimates, rows are different values of σ_Y. Within each graph, the y-axis varies β_U.

Figure 3 shows results for Scenarios (ii) and (iii) where we set $R_A = U$. Again we present results where $\gamma_A = 1, \gamma_U = 1, \gamma_{UA} = 0.5$ and $\sigma_A = 1$. Here α_U controls the strength of the relationship between A and U, and hence the extent of unmeasured confounding for scenario (iii), and $\alpha_U = 0$ is scenario (ii).

In this case, complete case analysis and model 3 (both regression and multiple imputation) estimate $\hat{\gamma}_A \approx 1$. This is unsurprising for complete case analysis, because it is the

conditional causal effect when $U = 0$, and there is no data available when $U = 1$. For all other approaches, the estimates depend on σ_Y. The estimates $\hat{\gamma}_R$ and $\hat{\gamma}_{RA}$ also depend on σ_Y; this is as expected based on the analytical result we presented earlier. With regard to MSE, we see again that regression imputation out-performs the other approaches; note that the complete case model's MSE is based on the complete data only, so not comparable.

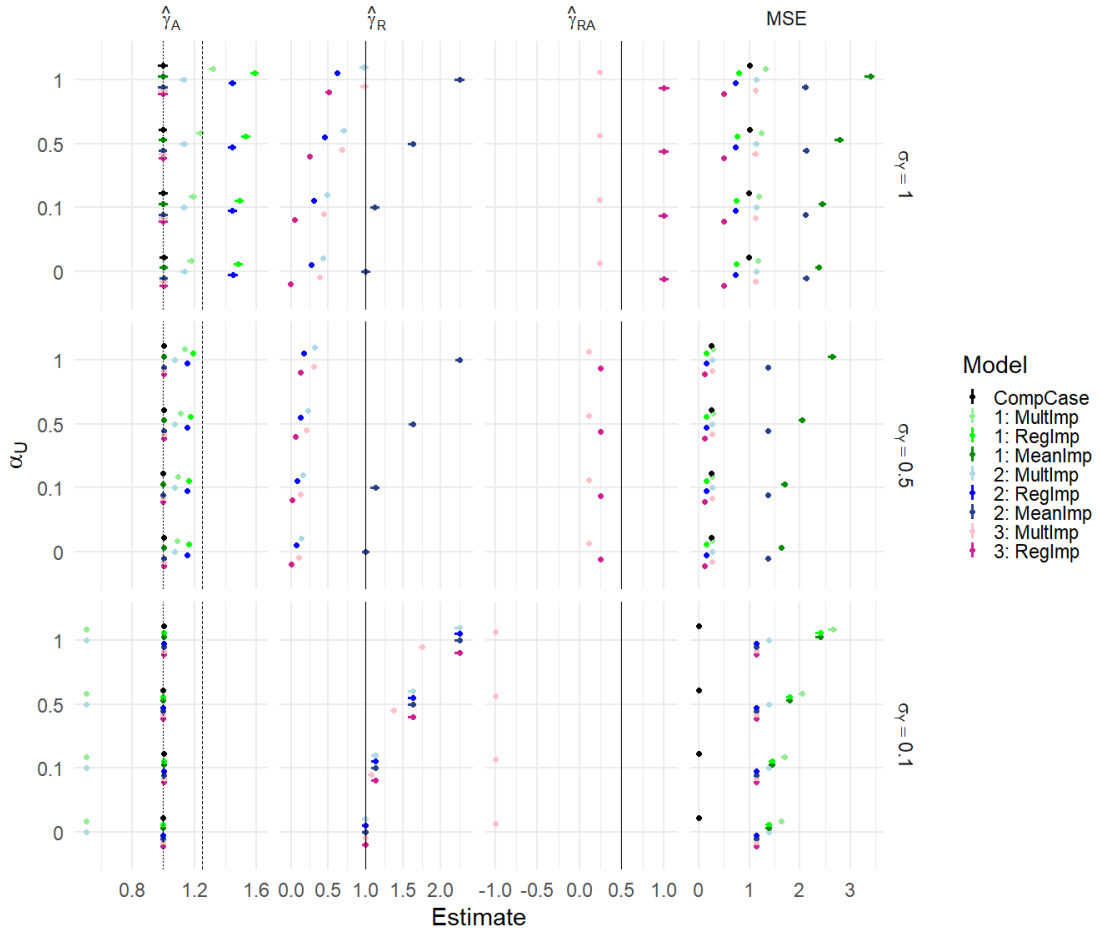

Figure 3: Results for scenarios (ii) and (iii) with $R_A = U$. Estimates with 95% empirical confidence intervals. Columns are different parameter estimates, rows are different values of σ_Y. Within each graph, the y-axis varies α_U.

Figure 4 shows results for Scenario (iii). In this case we have $\beta_A = \beta_{UA} = 0$ as defined by the scenario, and additionally we restrict to $\gamma_A = 1, \gamma_U = 1, \gamma_{UA} = 0.5, \sigma_A = 1$ and $\sigma_Y = 1$. The key parameters varying are α_U and β_U.

Increasing α_U changes the $\hat{\gamma}_A$ estimates. There is some, but lesser, impact of β_U. All imputation strategies yield similar results for $\hat{\gamma}_A$, except that bias is larger for regression imputation in models 1 and 2. However, β_U has a larger impact on the $\hat{\gamma}_R$ estimates. As expected, when β_U is larger (i.e. missingness is a stronger proxy for the unmeasured U) the $\hat{\gamma}_R$ estimate becomes larger, particularly for the multiple imputation models, although still much smaller than $\gamma_U = 1$. We again see that regression imputation has the smallest MSE, with model 3 regression imputation being the smallest.

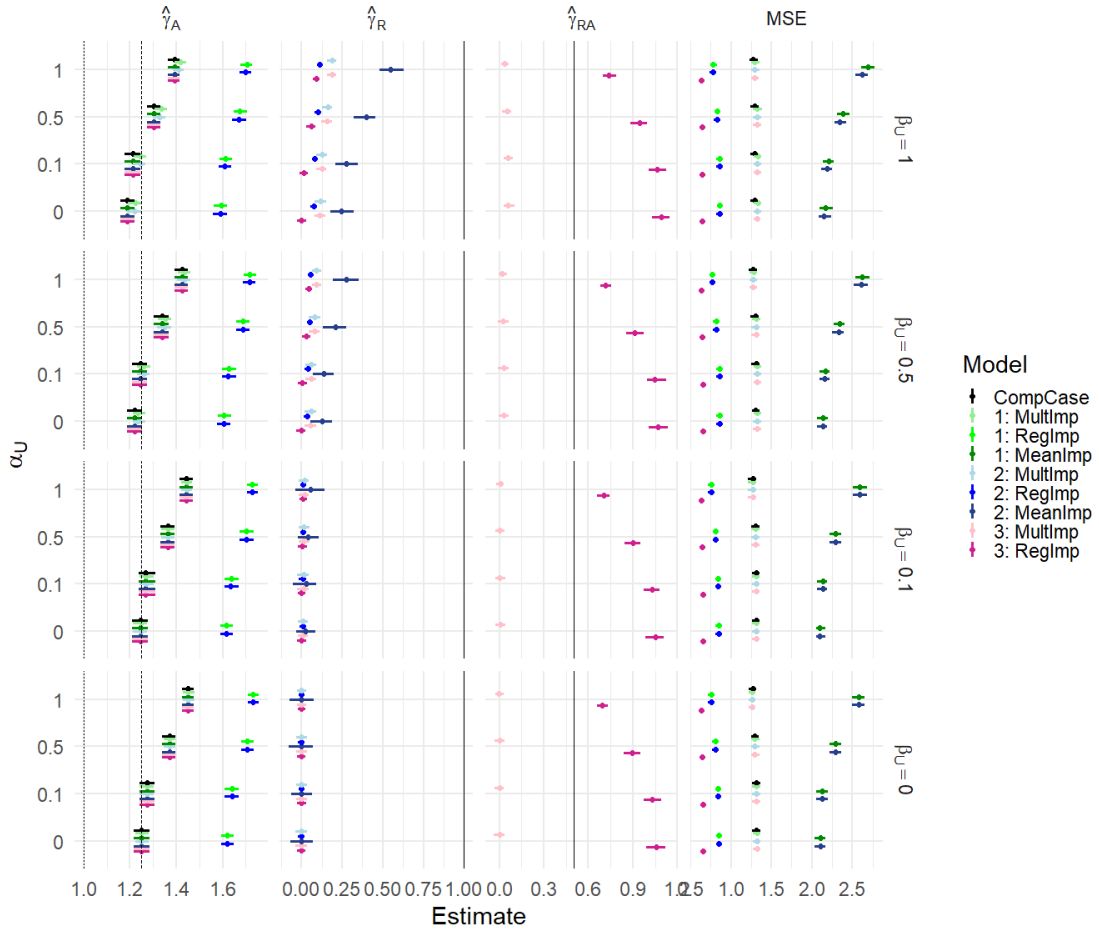

Figure 4: Results for scenario (iii). Estimates with 95% empirical confidence intervals. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies α_U.

Figure 5 shows results for Scenario (iv), where we examine the effects of missingness in A depending on A itself. Here, the scenario dictates that $\alpha_U = \beta_U = \beta_{UA} = 0$, and we additionally fix $\gamma_A = 1, \gamma_U = 1, \gamma_{UA} = 0.5$ and $\sigma_A = 1$. The key varying parameter is β_A which controls the dependence of R_A on A.

Here, complete case analysis and mean imputation provide unbiased estimates of the marginal causal effect of A on Y (1.25), as expected. Performing regression or multiple

imputation, then fitting outcome model 1 or 2 leads to bias. However, this is avoided when imputing and then fitting outcome model 3.

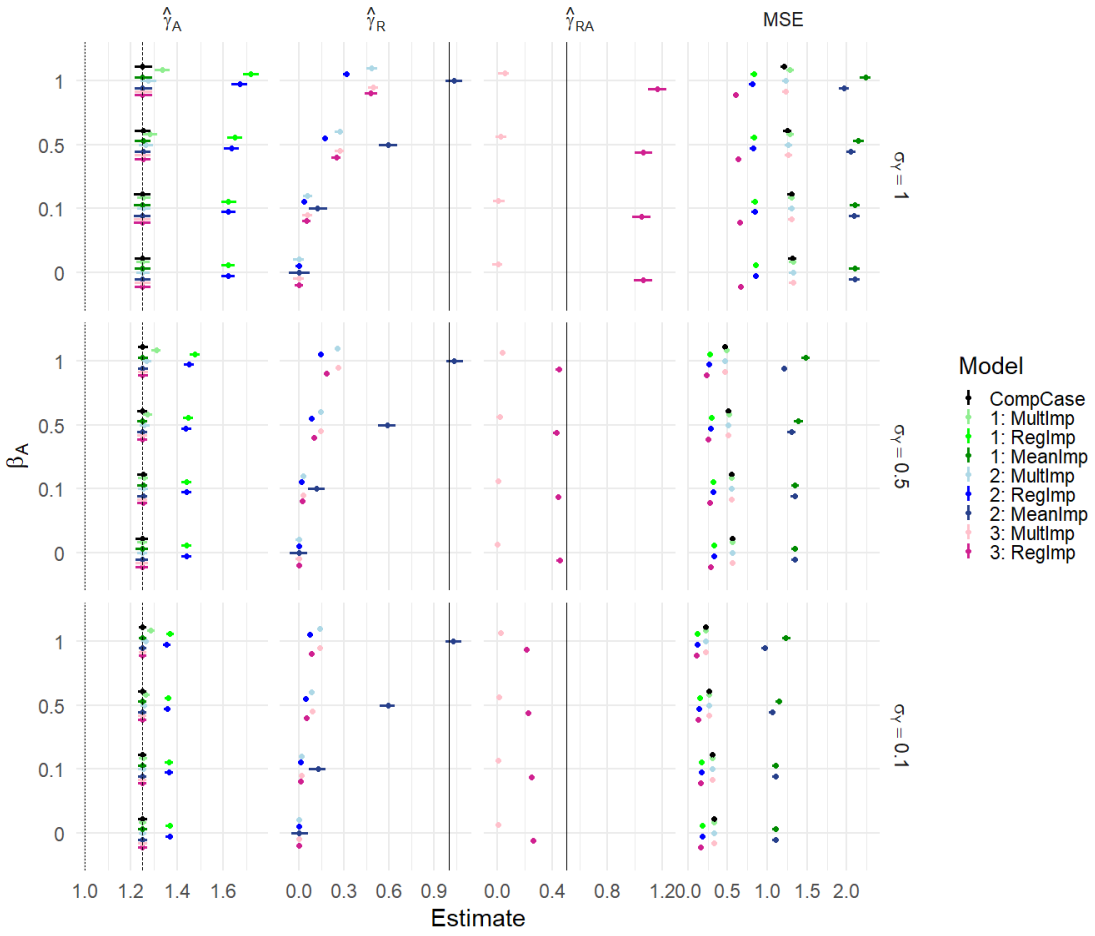

Figure 5: Results for scenarios (iv). Estimates with 95% empirical confidence intervals. Columns are different parameter estimates, rows are different values of σ_Y. Within each graph, the y-axis varies β_A.

Figure 6 shows results for Scenario (v). Here, the scenario dictates that $\alpha_U = 0$, and we additionally fix $\gamma_A = 1, \gamma_U = 1, \gamma_{UA} = 0.5, \sigma_A = 1, \sigma_Y = 1$ and $\beta_{UA} = 0$. The key varying parameters are β_A and β_U which control the dependence of R_A on A and on U respectively.

Now, when both β_U and β_A become large, all approaches are biased in estimating the marginal causal effect. The MIMI and complete case analyses underestimate γ_A while standard multiple and regression imputation tend to overestimate γ_A for models 1 and 2.

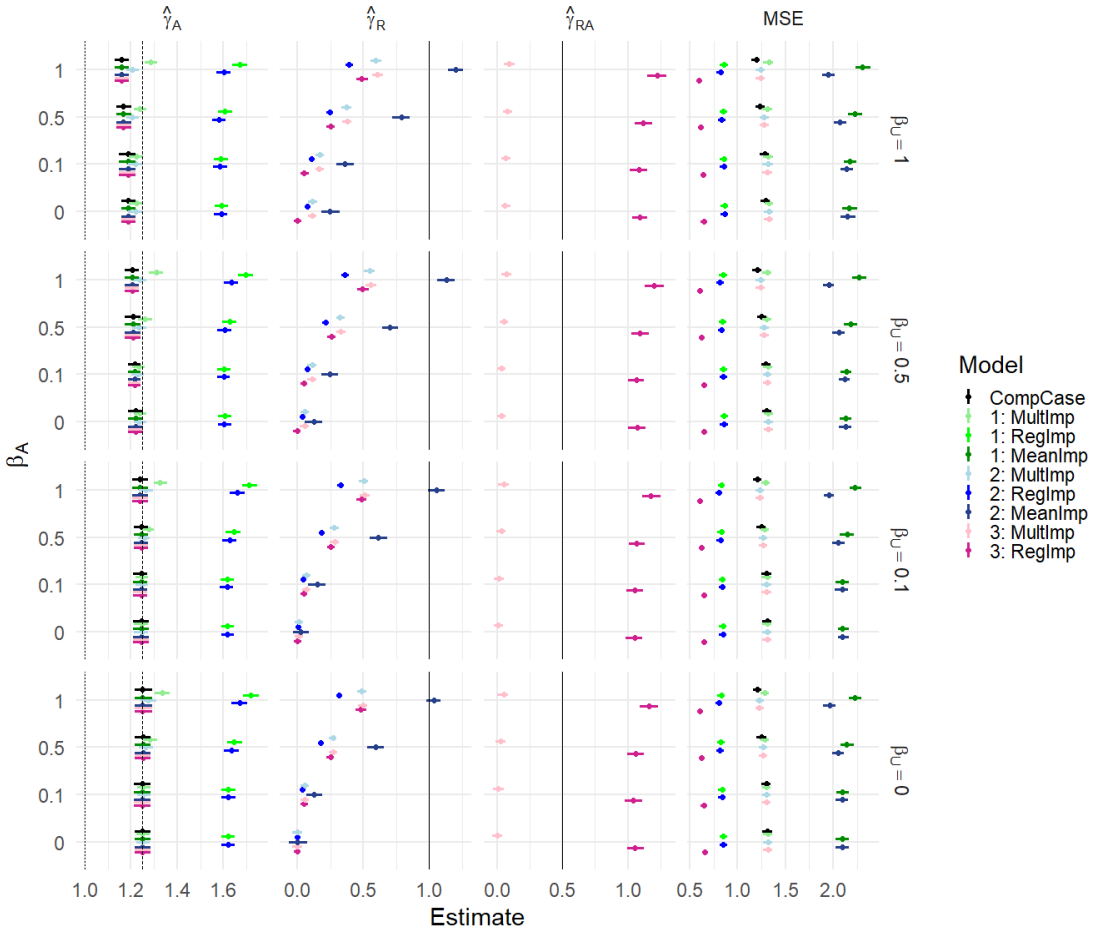

Figure 6: Results for scenarios (v). Estimates with 95% empirical confidence intervals. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies β_A.

Figure 7 shows 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.

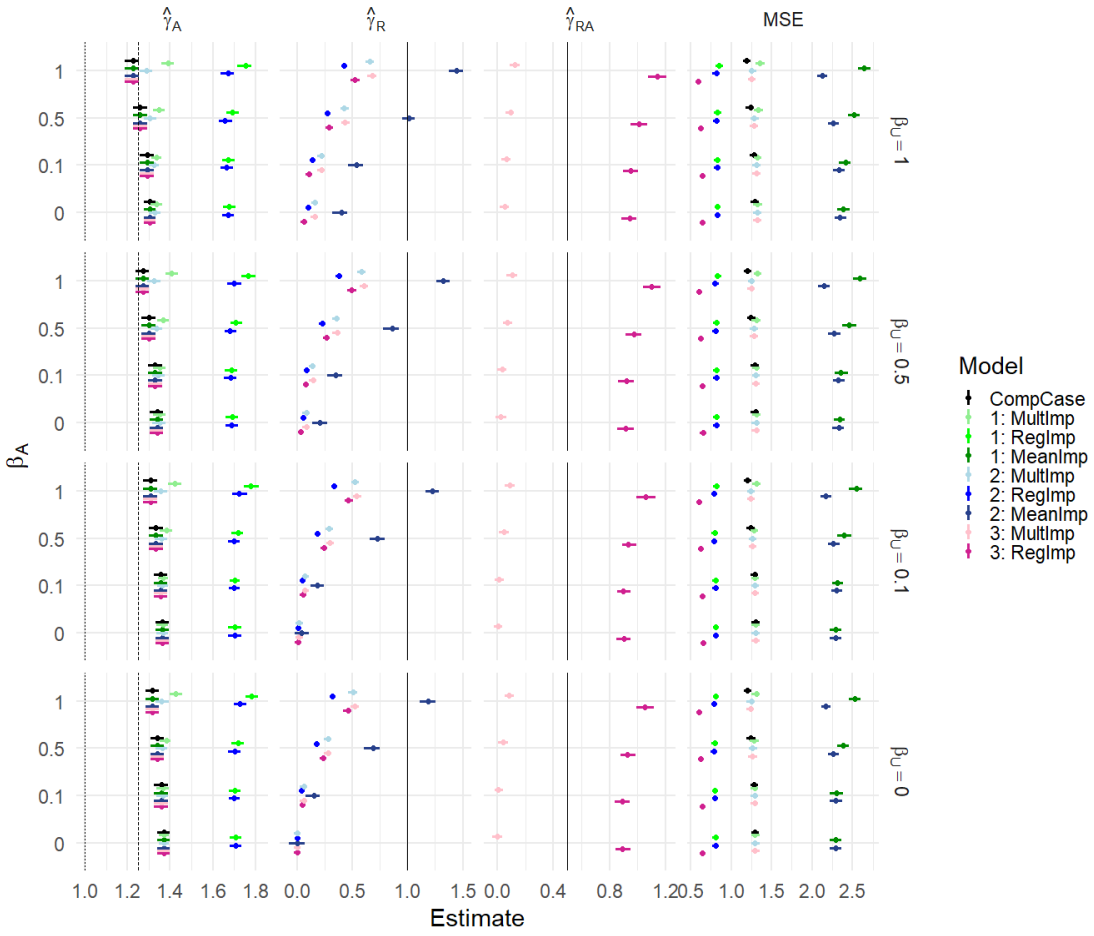

Figure 7: Results for scenario (vi). Estimates with 95% empirical confidence intervals. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies β_A.

The standard errors corresponding to all of the above Figures, both empirical and average model based, are given in Supplementary Figures 1 to 6. Across all scenarios, for $\hat{\gamma}_A$, the empirical and model-based standard errors were in broad agreement for multiple imputation and for complete case analysis. Model-based standard errors underestimated

the empirical standard error when using regression imputation, and overestimated when using unconditional mean imputation. This is all as expected.

For the standard error of $\widehat{\gamma_R}$, all model-based standard errors were conservative, except in the case of mean imputation where the empirical standard error exceeded the model-based standard error. For the standard error of $\widehat{\gamma_{RA}}$, model-based standard errors were conservative for multiple imputation, but underestimated the empirical standard error for regression imputation.

Discussion

In this paper we have explored, through simulation, the potential merits of supplementing a missing data strategy 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 or an unmeasured prognostic variable. We divide the main findings into implications for causal estimation, and implications for prediction.

Implications for causal estimation

In the MCAR scenario, without unmeasured confounding, adding a missing indicator was unlikely to introduce bias in estimation of causal effects. 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. Specifically, when unmeasured confounding exists, the missing indicator and/or its interaction with the main effect were estimated to be non-zero. Additionally, when missingness was perfectly

correlated with the unmeasured confounder, the measured effect was highly biased (see Appendix). Nevertheless, these non-zero effect estimates of the missing indicator act as a signal that it will be difficult or impossible to obtain unbiased causal effects. Alongside whether to incorporate missing indicators, we also explored the relative benefits of mean imputation, regression imputation and multiple imputation. As expected, between these approaches, multiple imputation was found to be the most robust. We found that in some MNAR scenarios where multiple imputation is usually biased, this bias is removed or alleviated by MIMI.

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 [4,5]. The idea of combining the missing indicator approach with multiple imputation was first proposed by [6], and has been further explored by [7] and [3]. In these 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, [3] in particular noted that the use of missing indicators can partly adjust for unmeasured confounding, similar to our findings.

Therefore, we recommend the use of MIMI (including interactions between missing indicators and the corresponding variable) as a strategy for handling missing data in causal estimation problems. Non-zero estimates of the missing indicator then alert to possible occurrence of MNAR, 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.

Implications for prediction

Regression imputation led to smaller MSE than corresponding multiple imputation approaches, especially when combined with a missing indicator and associated interaction term. Multiple imputation has long being assumed to be the best choice for handling missing data in prediction, despite the motivation for the approach coming from consideration of bias, which are only relevant for causal inference. Here we demonstrate that regression imputation (imputing the predicted mean rather than simulating from the posterior predictive distribution) leads to reduced MSE. This finding is likely due to the associated reduction in variance with little or no loss in information (since prediction focuses on prediction of Y, not causal estimation of parameters). However, care is needed in estimating the standard error of an associated predictive interval, since standard methods would underestimate this based on a single regression imputation. We also saw that regression imputation led to larger bias in effect estimates than other approaches. While such bias is not a direct concern in predictive modelling, causal effects are known to be more stable and robust over time and geography [13], and also allow for counterfactual prediction, which is useful in many decision support contexts [14,15].

For prediction specifically, [8] advocate the use of a pattern submodel, in which separate models are fit for each missingness pattern. Our Model 3 can be thought of as similar to this approach; as noted by [8], there is asymptotic equivalence between the approaches. The pattern submodel is easier to use in prediction, but hard to interpret from a causal effect

estimation perspective, so we did not consider it here. [16] also compared techniques to handle missing data in prediction; however, they did not consider the regression imputation technique that we found to be optimal in terms of MSE, and gave limited attention to MNAR mechanisms, which are likely in routine data where prediction models are commonly derived and applied.

Our prediction findings require further investigation, to ascertain whether the regression imputation strategy improves accuracy of a predictive model in real data. This should specifically be explored in the context of model calibration and discrimination. Nonetheless, it is worth noting that a further advantage of using regression imputation for prediction is that it is more feasible to apply at 'prediction time' – i.e. dealing with missing predictors when making a prediction for a new observation. Applying multiple imputation in this setting is often infeasible, and there are issues with using a different approach to imputation when developing a model, compared with the approach when the model is used in practice [8]. Therefore, we recommend that developers of predictive models consider regression imputation as an alternative approach to handle missing data, and base the choice of imputation method on the accuracy of the resulting models, and the feasibility of performing the required imputation at 'prediction time'.

Strengths and limitations

The paper has several strengths. 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, where

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.

Conclusions

We recommend that addition of a missing indicator, and corresponding interaction terms, can supplement, but not replace, the existing chosen imputation strategy. Where the goal is prediction, regression imputation should be explored as an alternative to multiple imputation, as this may increase both accuracy and practicality.

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

The authors have no specific funding to declare.

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 $R_A = U$

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 $\mathcal{J} = \{i: R_{A,i} = 0\}$, 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 [11], consider a hypothetical imputation model based on $\mu_{Y,i}$:

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

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 [11] (referencing [17]) 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,

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

Returning to 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 $\kappa = \frac{\gamma_A^* \sigma_A^{*2}}{\gamma_A^{*2} \sigma_A^{*2} + \sigma_Y^{*2}}$.

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 with

$$E_Y[\gamma_U^* - \gamma_A \gamma_U^* \kappa - \gamma_U] = 0.$$

Rearranging yields,

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

as claimed.

Supplementary figures

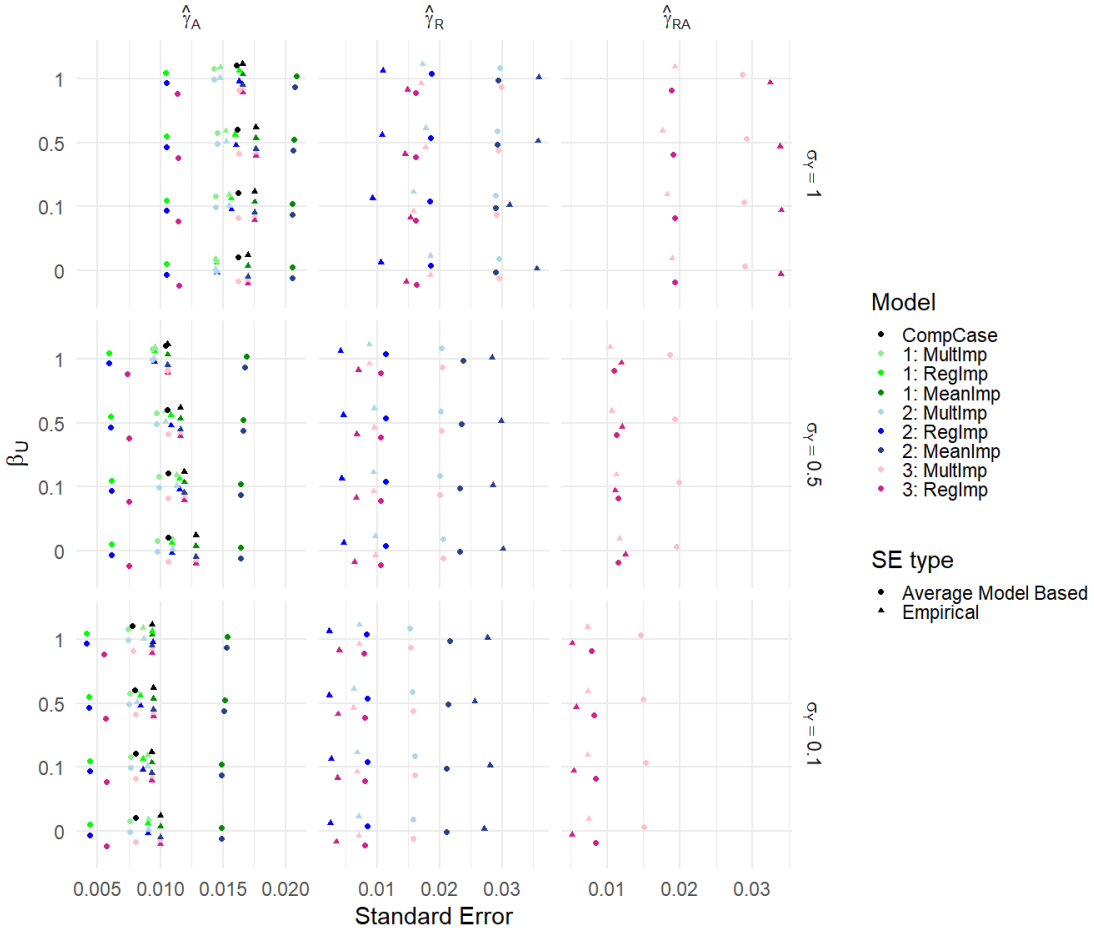

Figure S1: Results for scenarios (i) and (ii). Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of σ_Y. Within each graph, the y-axis varies β_U.

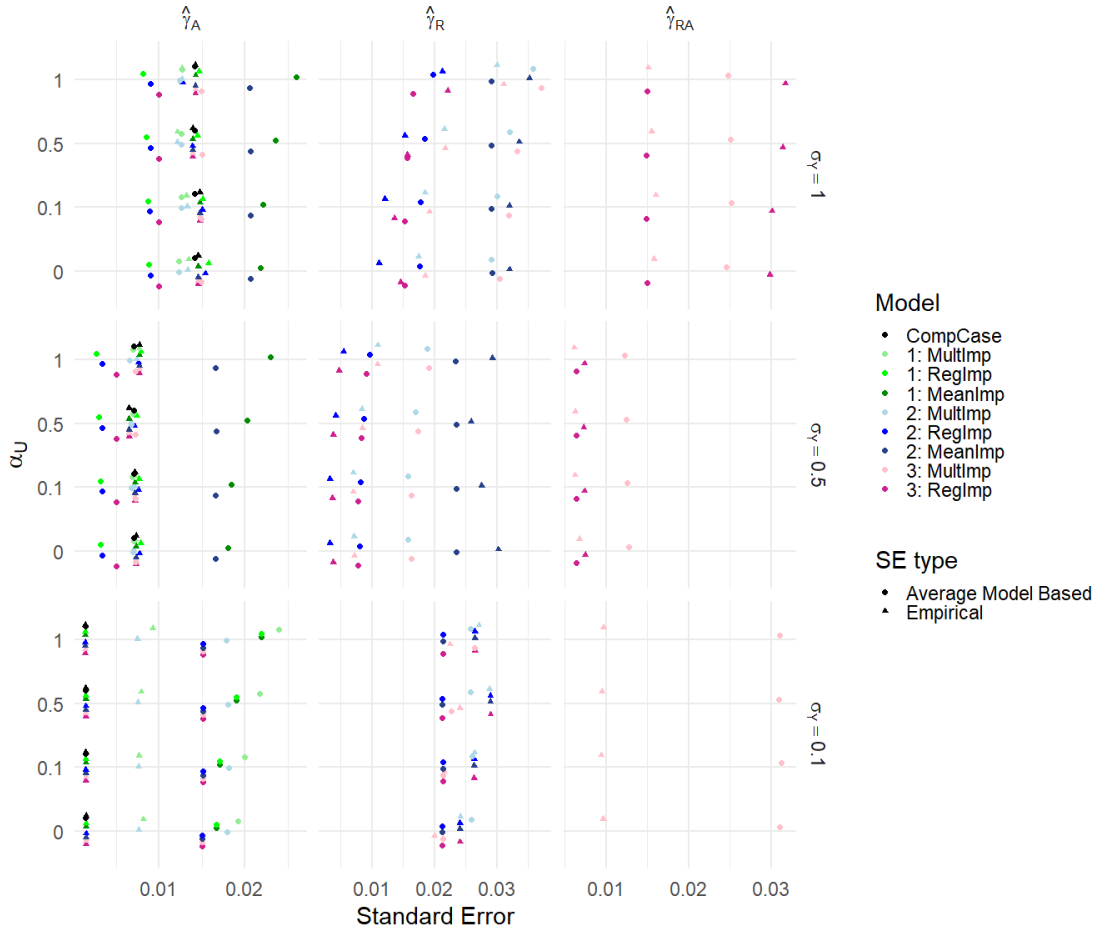

Figure S2: Results for scenarios (ii) and (iii) with $R_A = U$. Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of σ_Y. Within each graph, the y-axis varies α_U.

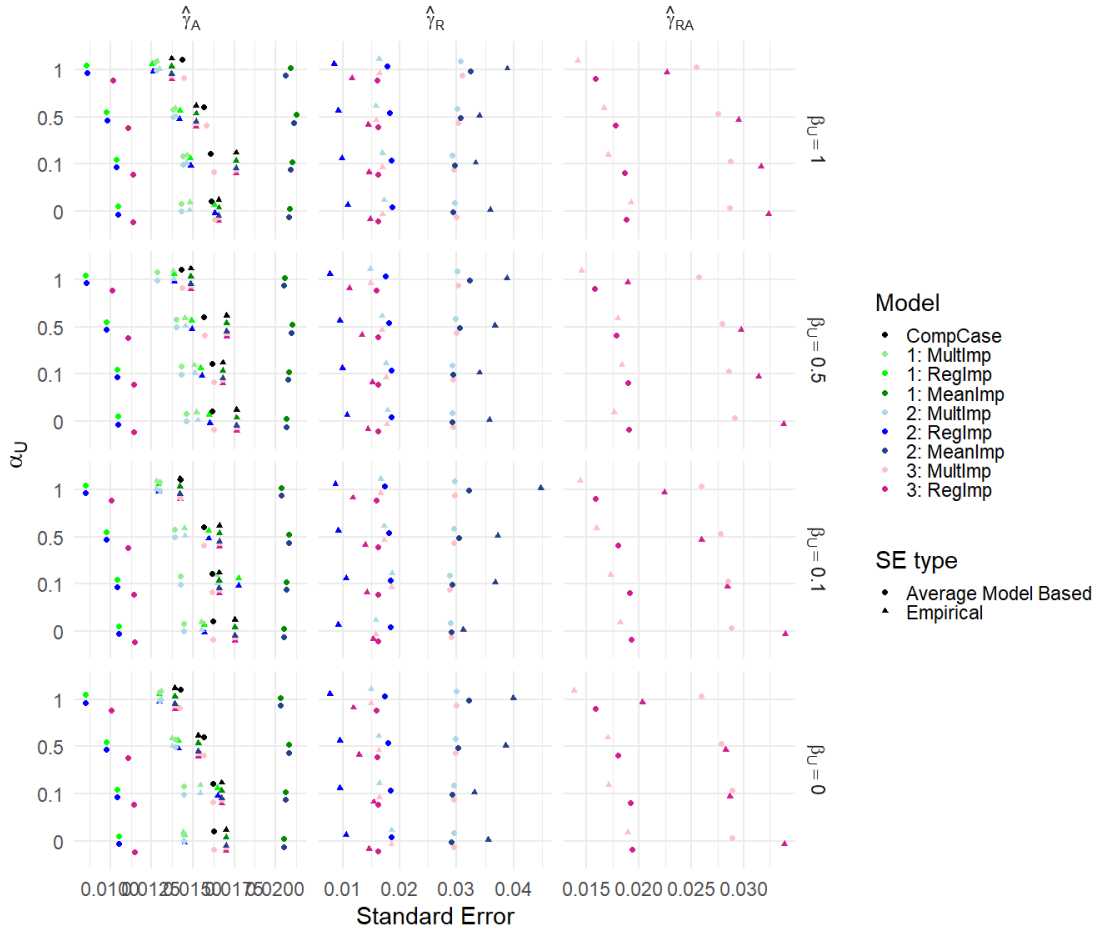

Figure S3: Results for scenario (iii). Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies α_U.

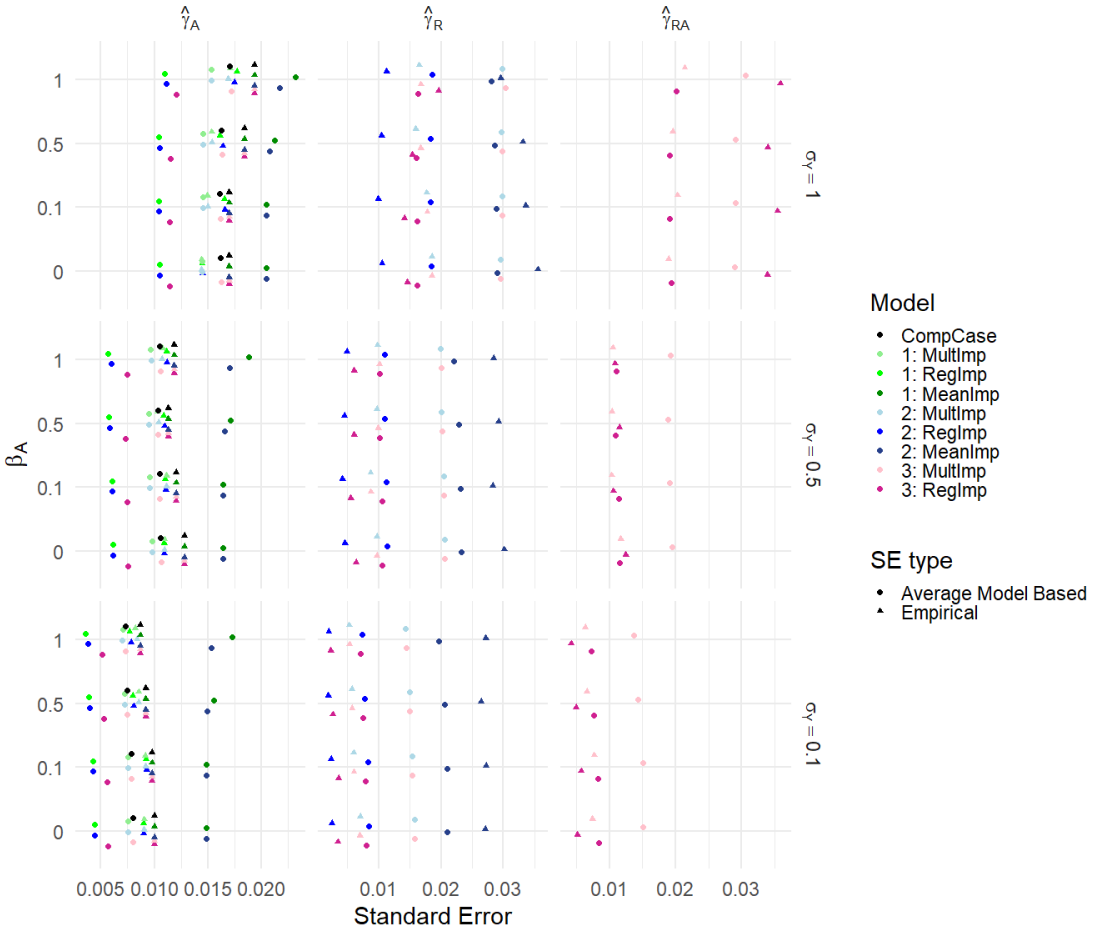

Figure S4: Results for scenarios (iv). Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of σ_Y. Within each graph, the y-axis varies β_A.

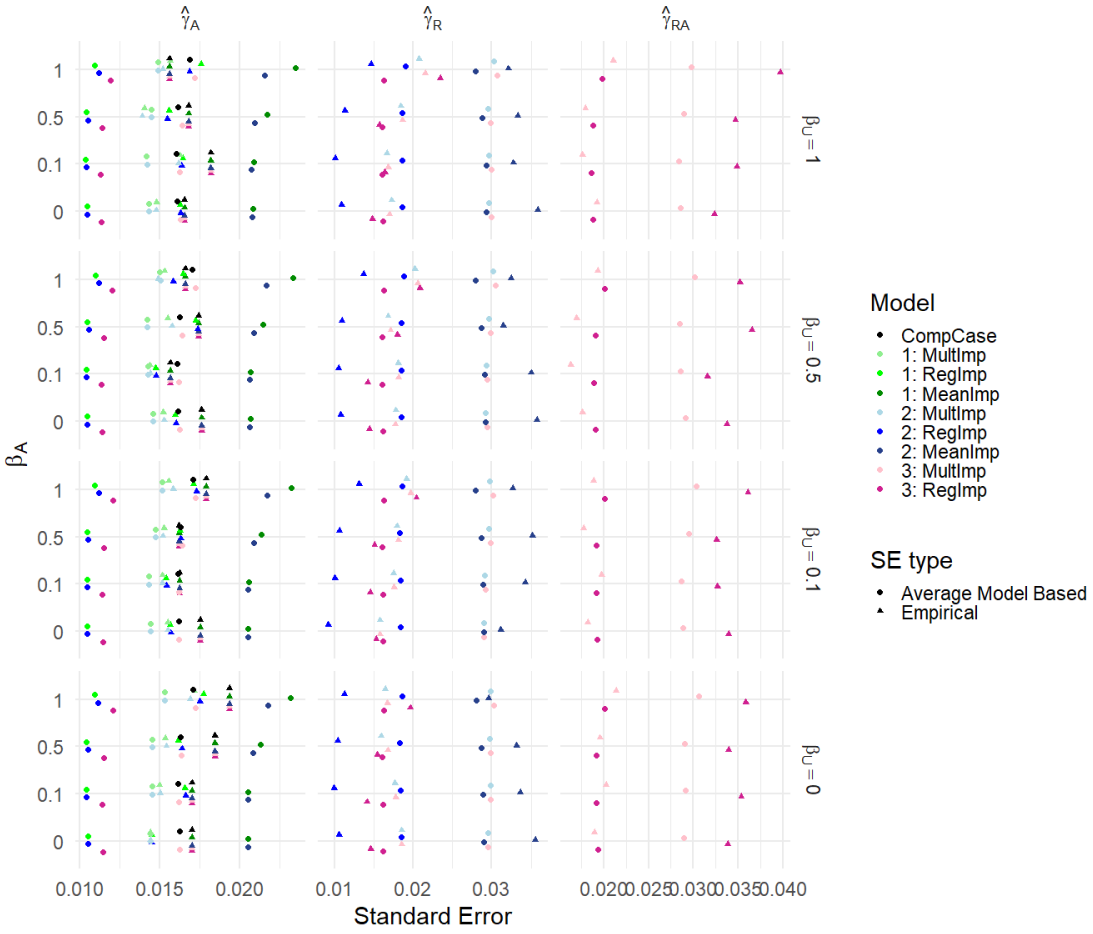

Figure S5: Results for scenarios (v). Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies β_A.

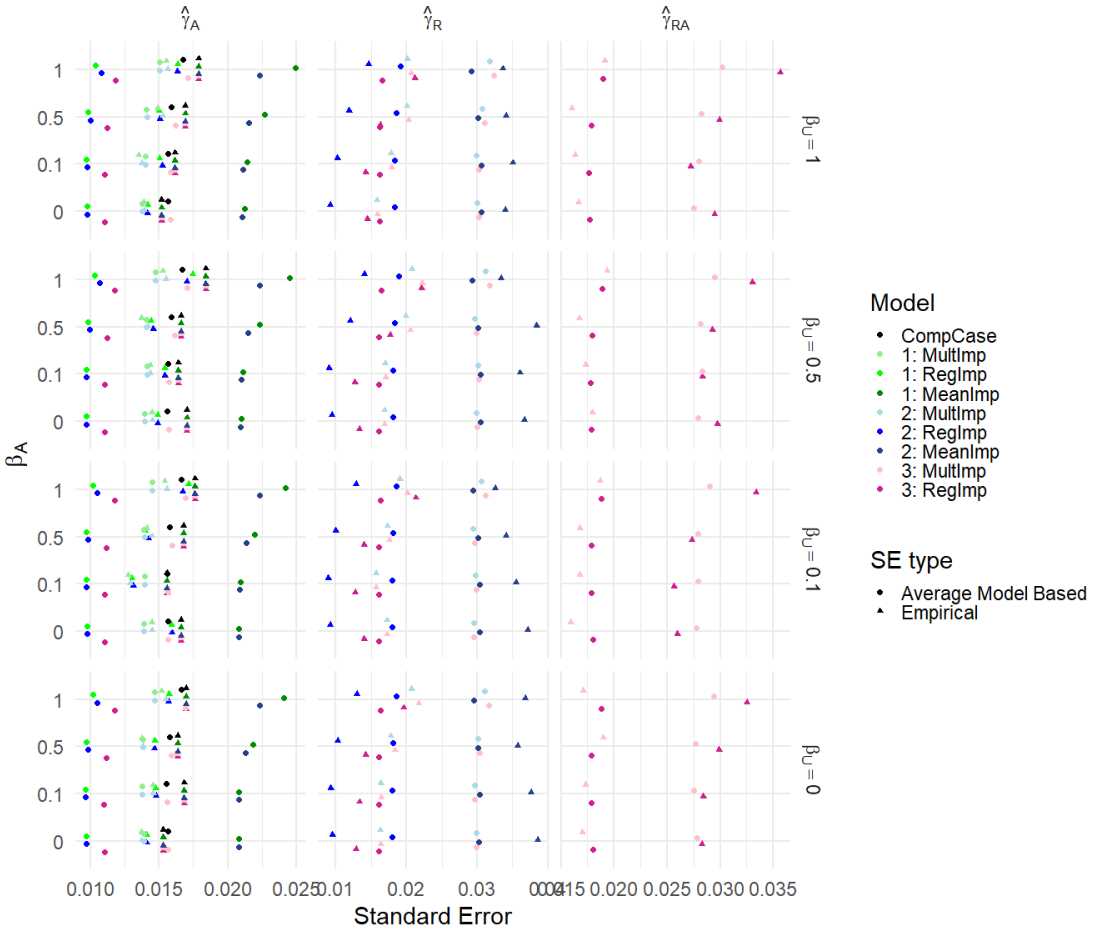

Figure S6: Results for scenario (vi). Average model based and empirical standard errors. Columns are different parameter estimates, rows are different values of β_U. Within each graph, the y-axis varies β_A.

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. *Advances in Neural Information Processing System* 2013;**26**:1–9. doi:10.1007/s15010-013-0464-5

3 Choi J, Dekkers OM, Cessie S le. A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology* 2019;**34**:23–36. doi:10.1007/s10654-018-0447-z

4 Knol MJ, Janssen KJ, Donders ART *et al.* Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology* 2010;**63**:728–36. doi:10.1016/j.jclinepi.2009.08.028

5 Groenwold RH, White IR, Donders ART *et al.* Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. *CMAJ* 2012;**184**:1265–9. doi:10.1503/cmaj.110977

6 Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in Medicine* 2009;**28**:1402–14. doi:10.1002/sim.3549

7 Seaman S, White I. Inverse Probability Weighting with Missing Predictors of Treatment Assignment or Missingness. *Communications in Statistics-Theory and Methods* 2014;**43**:3499–515. doi:10.1080/03610926.2012.700371

8 Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. *Biostatistics* Published Online First: September 2018. doi:10.1093/biostatistics/kxy040

9 Daniel RM, Kenward MG, Cousens SN *et al.* Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research* 2012;**21**:243–56. doi:10.1177/0962280210394469

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

11 Frost C, Thompson SG. Correcting for regression dilution bias: Comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society Series A: Statistics in Society* 2000;**163**:173–89. doi:10.1111/1467-985X.00164

12 Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019;**38**:2074–102. doi:10.1002/sim.8086

13 Steyerberg EW, Moons KG, Windt DA van der *et al.* Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS medicine* 2013;**10**:e1001381.

14 Hernán MA, Hsu J, Healy B. Data science is science's second chance to get causal inference right: A classification of data science tasks. *CHANCE* 2018;**32**:42–9. doi:10.1080/09332480.2019.1579578

15 Sperrin M, Martin GP, Pate A *et al.* Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Statistics in Medicine* 2018;**37**:4142–54. doi:10.1002/sim.7913

16 Marshall A, Altman DG, Royston P *et al.* Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Medical Research Methodology* 2010;**10**:7. doi:10.1186/1471-2288-10-7

17 Snedecor GW, Cochran WG. Statistical Methods. 6th Edition. *Applied Statistics* Published Online First: 1968. doi:10.2307/2985653