Multiple imputation (MI) is increasingly being used to handle missing data in epidemiologic research. When data on both the exposure and the outcome are missing, an alternative to standard MI is the “multiple imputation, then deletion” (MID) method, which involves deleting imputed outcomes prior to analysis. While MID has been shown to provide efficiency gains over standard MI when analysis and imputation models are the same, the performance of MID in the presence of auxiliary variables for the incomplete outcome is not well understood. Using simulated data, we evaluated the performance of standard MI and MID in regression settings where data were missing on both the outcome and the exposure and where an auxiliary variable associated with the incomplete outcome was included in the imputation model. When the auxiliary variable was unrelated to missingness in the outcome, both standard MI and MID produced negligible bias when estimating regression parameters, with standard MI being more efficient in most settings. However, when the auxiliary variable was also associated with missingness in the outcome, alarmingly MID produced markedly biased parameter estimates. On the basis of these results, we recommend that researchers use standard MI rather than MID in the presence of auxiliary variables associated with an incomplete outcome.

Abbreviations: MAR, missing at random; MI, multiple imputation; MID, multiple imputation, then deletion.
an imputation model is generated including \( Y \) and all components of \( \mathbf{X} \). To ensure that imputation and analysis models are consistent and to avoid biasing associations towards independence, observed and imputed values of \( Y \) are used to impute missing values for all components of \( \mathbf{X} \) and vice versa (17–19). Following imputation, in standard MI all of the observed and imputed data for \( Y \) and \( \mathbf{X} \) are used in the analysis of each of the completed data sets. In contrast, MID excludes (or deletes) cases with imputed \( Y \)'s from the analysis of each of the completed data sets. In other words, analysis using standard MI involves all participants in the study, whereas analysis under MID is restricted to participants with observed outcome data. Provided that the MAR assumption is valid, the deletion of observations with imputed outcomes in MID offers 2 practical advantages. Firstly, for a finite number of imputations, MID has been shown to produce more precise estimates of \( \theta_{Y|X} \) than standard MI (i.e., smaller standard errors, narrower confidence intervals), although efficiency gains tend to be minor unless the number of imputations is small and the proportion of missing data is high (8). Secondly, removing observations with imputed outcomes from the analysis can help to minimize the bias introduced by a misspecified model for imputing the missing outcomes (8).

The rationale behind MID is that following imputation, cases with missing outcome data do not contribute any further information about the parameters \( \theta_{Y|X} \); hence, retaining these cases in the analysis only adds noise to the estimation process (7, 8). While this assertion is correct when the imputation and analysis models include the same variables (in an appropriate form), in practice these models often differ. Indeed, one of the appealing features of the MI framework is the ability to incorporate additional “auxiliary” variables into the imputation model that are not part of the substantive analysis to improve the prediction of missing values (14). In clinical trials, for example, postrandomization measures such as treatment compliance are often used as auxiliary variables. Importantly, while both standard MI and MID benefit equally from the inclusion of auxiliary variables to improve the prediction of missing values in \( \mathbf{X} \), only standard MI benefits from the inclusion of auxiliary variables to predict missing values in \( Y \) (8). However, depending on the number of imputations used, the additional information provided by an auxiliary variable for \( Y \) needs to be fairly substantial for standard MI to demonstrate efficiency advantages over MID. On the basis of a simulation study involving normally distributed variables, von Hippel found that MID was more efficient than standard MI when the correlation between a single auxiliary variable and the incomplete outcome did not exceed 0.7, 0.6, and 0.5 for 2, 5, and 10 imputations, respectively (8). It is unclear whether MID would maintain similar efficiency advantages over standard MI with a larger number of imputations.

While important, efficiency gains are not the only consideration when identifying auxiliary variables for inclusion in imputation models. Arguably the more essential role of auxiliary variables is in helping to make the MAR assumption which underlies MI more plausible. In developing high-quality imputations, numerous experts have recommended the inclusion in the imputation model of auxiliary variables that are associated with the incomplete variables to be imputed, the probability of missing data, or both (e.g., 20–24). It is the inclusion of auxiliary variables related to the probability of missing data that is important for satisfying the MAR assumption. Auxiliary variables related to the probability of missing data were not considered in von Hippel’s original paper proposing MID (8). In a landmark study, Collins et al. (22) demonstrated via simulation that failure to incorporate information from auxiliary variables that are correlated with an incomplete outcome and with missingness in the outcome leads to biased inference in estimating regression coefficients from linear regression models following MI. Given the potential for auxiliary variables to reduce bias and improve efficiency, they recommended that researchers adopt inclusive strategies for selecting auxiliary variables to include in imputation models. These findings have important implications for the use of MID in studies where auxiliary information is available. Since MID is unable to take advantage of auxiliary information for an incomplete outcome, it can be argued that the approach is not entirely consistent with the inclusive strategy for variable selection when setting up an imputation model. Further, if auxiliary variables are required to satisfy a MAR assumption for the outcome, it is unclear whether including these variables in the imputation model and then deleting imputed outcomes prior to analysis could introduce bias. To our knowledge, these issues have not been investigated in the comparison of standard MI and MID.

Our aim in this paper was to evaluate the performance of standard MI and MID in regression settings where data are missing for both the outcome and the exposure and where auxiliary variables associated with the outcome are included in the imputation model. We hypothesized that the efficiency advantages of MID would be less pronounced with a larger number of imputations, and that this approach would introduce bias in the estimation of \( \theta_{Y|X} \) when the imputation model contained auxiliary variables that were additionally associated with the probability of missing data on the outcome.

**METHODS**

**Simulation study**

We evaluated the performance of standard MI and MID in the presence of an auxiliary variable associated with an incomplete outcome by extending the earlier simulation study of von Hippel (8). Using the same data generation procedure, we investigated the consequences of using a larger number of imputations and allowing for missingness in the outcome to depend on an auxiliary variable.

For each simulation scenario, 1,000 complete data sets of size \( n = 200 \) were created. Initially, 2 predictor variables \( X_1 \) and \( X_2 \) were generated from a bivariate standard normal distribution with correlation \( \rho_{12} \). An outcome \( Y \) was then produced according to the linear regression model \( Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon \), where \( \epsilon \) was a normally distributed error term with mean 0 and variance \( \sigma^2 \), and where the regression parameters (\( \alpha, \beta_1, \beta_2 \)) were set to (1, 1, 1). The proportion of the variance in \( Y \) explained by the linear regression model \( (R^2) \) was fixed by setting the variance as \( \sigma^2 = 2(1 - R^2)(1 + \rho_{12})/R^2 \). Next, a standard normal auxiliary variable \( Z \) was generated according to the
equation \( Z = u + \rho_{xy}Y / \text{var}(Y) \), where \( \rho_{xy} \) was the correlation between \( Z \) and \( Y \) and where \( u \) was normally distributed with mean 0 and variance \( 1 - \rho_{xy}^2 \). In generating complete data sets, \( R^2, \rho_{12}, \) and \( \rho_{xy} \) were independently varied. Following the simulation study of von Hippel (8), we allowed \( \rho_{12} \) and \( R^2 \) to take the values 0.2, 0.5, and 0.8, while \( \rho_{xy} \) was set to either 0.1, 0.5, or 0.9. Collectively this resulted in 27 scenarios with complete data to investigate.

Following the generation of complete data sets, values of \( X_2 \) and \( Y \) were independently set to missing according to one of 2 MAR mechanisms. In one setting, we replicated the “coordinated missingness” mechanism previously considered by von Hippel in which \( X_2 \) and \( Y \) were set to missing independently with probability \( 2p(1 - \Phi(X_2)) \), where \( \Phi \) is the cumulative distribution function of the standard normal distribution. Hereafter we refer to this missing-data mechanism as “auxiliary independent missingness,” since missingness in the outcome is conditionally independent of the auxiliary variable \( Z \). The motivation for investigating this missing-data mechanism was to evaluate the efficiency of standard MI and MID when a larger number of imputations was used; only 2, 5, and 10 imputations were considered previously. In a second setting, we considered a new missing-data mechanism in which values of \( Y \) were set to missing with probability \( 2p(1 - \Phi(X_2 + Z)) / \text{var}(X_2 + Z) \). \( X_2 \) was again set to missing with probability \( 2p\Phi(X_2) \). Throughout the remainder of the paper, we refer to this second missing-data mechanism as “auxiliary dependent missingness.” When setting values to missing, we allowed the overall proportion \( \rho \) of missing data in both \( X_2 \) and \( Y \) to equal 0.2 or 0.5. Together this resulted in 4 missing data patterns and 108 simulation scenarios overall.

### Imputation and analysis methods

For each simulation scenario, missing values in \( Y \) and \( X_2 \) were imputed using a Markov chain Monte Carlo algorithm assuming multivariate normality (20). \( Y, X_1, X_2, \) and \( Z \) were all included in the imputation model. Under auxiliary independent missingness, the expected percentage of incomplete cases was 34.7% and 66.7% when the proportion of missing data in \( X_2 \) and \( Y \) was equal to 0.2 and 0.5, respectively. Based on the rule of thumb that Monte Carlo error should be acceptably small when the number of imputations equals the percentage of incomplete cases (23), the use of approximately 70 imputations is recommended for standard MI. However, since the efficiency advantages of MID are greater when the number of imputations is lower (8) and since fewer imputations are common in practice, we chose 50 imputations as a reasonable compromise. Following imputation, the 50 complete data sets were analyzed directly for standard MI and analyzed following the deletion of observations with imputed outcomes for MID. Thus, for each scenario, standard MI and MID estimates were based on the same underlying imputed data. Each imputed data set was analyzed by fitting a linear regression model of the form \( Y = \alpha + \beta_1X_1 + \beta_2X_2 + \epsilon \). Of interest were the standard MI and MID estimates and 95% confidence intervals for the parameters \( \alpha, \beta_1, \) and \( \beta_2 \). Inference on individual parameters was obtained by combining estimates over the 50 imputed data sets using Rubin’s rules (3).

### Comparisons

For each simulation scenario, standard MI and MID parameter estimates across the 1,000 simulated data sets were summarized. The performance of the 2 approaches was assessed in terms of the bias (defined as the average difference between the parameter estimate and the true underlying value used to generate the data \( (\alpha = \beta_1 = \beta_2 = 1) \)) and the average estimated standard error of the parameter estimates. We also report the coverage of the estimated 95% confidence intervals, defined as the proportion of 95% confidence intervals that contained the true value. Based on 1,000 simulated data sets and a normal approximation to the binomial distribution, on 95% of occasions we would expect the coverage to lie between 0.936 and 0.964 for a nominal level of 0.95. In addition to summaries for each individual simulation scenario, mean values for the bias, average standard error, and coverage were also calculated across simulation scenarios for the 2 missing-data mechanisms to obtain an overall measure of performance.

All statistical calculations were performed using SAS, version 9.3 (SAS Institute, Inc., Cary, North Carolina). Multiple imputation was carried out using the MI procedure, while analysis was performed using the GENMOD and MIANALYZE procedures. Starting seeds for generating variables, inducing missing data, and performing MI were varied across simulation scenarios and recorded so that results could be reproduced.

### Binary variables

To investigate whether the performance of MID depends on variable type, we also performed a limited simulation study involving a binary outcome, a binary auxiliary variable,

---

**Table 1. Mean Values for Performance Measures Across 54 Scenarios Where Missing Data Were Induced Under the Auxiliary Independent Mechanism**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Standard MI</th>
<th>MID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias*</td>
<td>Range</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>0.001</td>
<td>-0.016 to 0.020</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.005</td>
<td>-0.023 to 0.042</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-0.006</td>
<td>-0.054 to 0.023</td>
</tr>
</tbody>
</table>

Abbreviations: MI, multiple imputation; MID, multiple imputation, then deletion; SE, standard error.

* Monte Carlo error for bias in \((\alpha, \beta_1, \beta_2) \leq (0.015, 0.025, 0.025)\) for standard MI and MID across the 54 scenarios.
and 2 binary covariates. Details of this additional simulation study are outlined in the Web Appendix (available at http://aje.oxfordjournals.org/).

RESULTS

Table 1 summarizes the performance of standard MI and MID under the auxiliary independent mechanism. Across the 54 simulation scenarios, both standard MI and MID exhibited negligible bias (i.e., the ranges of biases were consistent with Monte Carlo error), with coverage probabilities close to nominal levels throughout. In most settings, standard MI demonstrated moderate efficiency advantages over MID, with overall average standard errors (i.e., averaged across the 54 scenarios × 1,000 data sets) for the estimated parameters $\alpha$, $\beta_1$, and $\beta_2$ being at least 3% smaller with standard MI.

The efficiency advantages of standard MI under the auxiliary independent mechanism depended most strongly on the correlation between the auxiliary variable and the outcome (in the regression model $\rho_{zy}$) and on the proportion of missing values in the outcome variable ($\rho_{yz}$). Table 2 compares the performance of the 2 imputation approaches for different values of $\rho_{xy}$ when $\rho_{12} = 0.2$, $R^2 = 0.2$, and $P = 0.5$. For $\rho_{zy} = 0.1$, the average estimated standard errors across the 1,000 imputations for the 3 parameters were approximately 1% larger using standard MI compared with MID. When the correlation $\rho_{zy}$ was increased to 0.5, standard MI began exhibiting efficiency advantages over MID, particularly in estimating $\alpha$ and $\beta_1$. In this setting, the average estimated standard errors for $\alpha$ and $\beta_1$ were approximately 6% smaller using MI, and they were 2% smaller for $\beta_2$. Finally, for $\rho_{zy} = 0.9$, the average estimated standard errors were noticeably reduced with standard MI. Compared with MID, standard errors for $\alpha$, $\beta_1$, and $\beta_2$ were 28%, 27%, and 13% smaller using standard MI, respectively. A similar pattern of results was observed when the proportion of missing values in $Y$ and $X_2$ was 0.2; however, absolute differences in precision were less pronounced (results not shown).

As demonstrated in Table 3, standard MI also performed well under the auxiliary dependent mechanism. The absolute bias of standard MI was at most 0.023 across the 54 simulation scenarios for all 3 parameters, and the coverage probabilities remained close to nominal levels throughout. In contrast, MID showed deficiencies when the probability of missing data in the outcome variable depended on the auxiliary variable. The average bias and coverage for ($\alpha$, $\beta_1$, $\beta_2$) across the 54 simulation scenarios were ($-0.207$, $-0.074$, $-0.017$) and ($0.812$, $0.928$, $0.947$), respectively. The performance of MID suffered most when the proportion of missing data in $Y$ and $X_2$ was high (0.5), when the correlation between the auxiliary variable and the outcome was high (0.9), and when the proportion of variance in $Y$ explained by the regression model was low (0.2). Table 4 shows the performance of standard

<table>
<thead>
<tr>
<th>$\rho_{zy}$</th>
<th>$\rho_{12}$</th>
<th>$R^2$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 3. Mean Values for Performance Measures Across 54 Scenarios Where Missing Data Were Induced Under the Auxiliary Dependent Mechanism

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bias*</th>
<th>SE</th>
<th>Coverage</th>
<th>Bias*</th>
<th>SE</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>-0.001</td>
<td>0.019</td>
<td>0.948</td>
<td>0.937 to 0.962</td>
<td>-0.020</td>
<td>0.022</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.002</td>
<td>0.255</td>
<td>0.947</td>
<td>0.932 to 0.962</td>
<td>-0.074</td>
<td>0.028</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.004</td>
<td>0.264</td>
<td>0.945</td>
<td>0.933 to 0.957</td>
<td>-0.017</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Abbreviations: MI, multiple imputation; MID, multiple imputation, then deletion; SE, standard error.

* Monte Carlo error for bias in ($\alpha$, $\beta_1$, $\beta_2$) ≤ (0.015, 0.025, 0.026) for standard MI and MID across the 54 scenarios.

MI and MID under auxiliary dependent missingness for different values of $\rho_{zy}$ when $\rho_{12} = 0.2$, $R^2 = 0.2$, and $P = 0.5$. The bias associated with MID was relatively small when $\rho_{zy} = 0.1$, although there was some evidence of undercoverage in the estimation of $\alpha$ and $\beta_2$. For $\rho_{zy} = 0.5$, the bias in MID estimates was larger, particularly for $\alpha$ and $\beta_1$. Finally, for $\rho_{zy} = 0.9$, MID produced substantially biased estimates for $\alpha$ and $\beta_1$, with coverage dropping to just 0.114 for $\alpha$.

To more accurately demonstrate the bias introduced by MID in the presence of an auxiliary variable associated with the outcome and with missingness in the outcome, we performed additional simulations for $\rho_{12} = 0.2$ and $R^2 = 0.2$, where we varied the correlation between the auxiliary variable and the outcome ($\rho_{zy}$) in increments of 0.1. The performance of standard MI and MID in estimating $\alpha$ and $\beta_1$ are plotted in Figure 1. As shown in Figure 1A, estimates of $\beta_1$ were close to the true value for both standard MI and MID when the proportion of missing data in $Y$ and $X_2$ was 0.2. However, when the proportion of missing data in $Y$ and $X_2$ was increased to 0.5, MID exhibited bias, even for small values of $\rho_{zy}$, with the magnitude of the bias increasing linearly with the correlation $\rho_{zy}$. A similar pattern of results was observed for $\alpha$ (Figure 1B), although for this parameter MID also exhibited some bias when the proportion of missing data in $Y$ and $X_2$ was 0.2.

In line with results for continuous outcomes, standard MI performed well when missing data in a binary outcome depended on an auxiliary variable, but coefficient estimates in a logistic regression model were biased with MID (see Web Table 1). Once again the magnitude of the bias of MID depended on the strength of the association between the outcome and the auxiliary variable.

**DISCUSSION**

In this study, we evaluated the performance of standard MI and MID when the imputation model was enriched by auxiliary information for the incomplete outcome. In line with

![Figure 1](https://example.com/f1.png)
previous results, both standard MI and MID exhibited negligible bias in estimating regression parameters when an auxiliary variable associated with the incomplete outcome, but not with missingness in the outcome, was added to the imputation model. We have now demonstrated that when the auxiliary variable is also related to missingness in the outcome and hence is required in the imputation model to satisfy the MAR assumption, MID produces biased estimates of regression parameters, whereas standard MI does not. These results have important implications for the use of MID in applied research.

When the auxiliary variable was unrelated to missingness in the outcome, results demonstrated that the precision of MID was only marginally better than that of standard MI for a weak correlation between the auxiliary variable and the outcome. Conversely, standard MI was noticeably more efficient for moderate-to-strong correlations between the auxiliary variable and the outcome. The results are in line with those observed previously for 10 or fewer imputations (8); however, in our study, the efficiency advantages of standard MI were greater with 50 imputations. This suggests that the intended number of imputations is an important factor to take into account when choosing between standard MI and MID based solely on efficiency considerations. Although early texts on MI suggested that 10 or fewer imputations are usually adequate (19, 25, 26), more recent recommendations state that the number of imputations should be much larger (i.e., 20–100) (23, 27). Since increasing the number of imputations entails greater precision, standard MI with a large number of imputations should be preferred over MID if the primary goal is to maximize efficiency. In light of continuing improvements in computational power and analytical software, standard MI with a large number of imputations should be feasible in most practical settings.

When missingness in the outcome depended on the auxiliary variable, MID produced biased estimates of regression parameters, with the magnitude of the bias being positively associated with the amount of missing data and the correlation between the auxiliary variable and the outcome. Effectively, MID discarded the information about the outcome provided by the auxiliary variable, leading to violation of a MAR assumption that was otherwise satisfied under standard MI. The results suggest that MID is not an optimal strategy in the presence of auxiliary variables that are associated with missingness in the outcome. In our view, failing to exploit the information offered by auxiliary variables and potentially introducing serious bias into the analysis for small potential gains (or possible losses) in precision is a poor trade-off. This leaves researchers with 2 choices for implementing MI when auxiliary information for an incomplete outcome is available: 1) imputing using a model that excludes auxiliary variables associated with the incomplete outcome and proceeding with MID or 2) incorporating these auxiliary variables into the imputation model and employing a standard MI analysis. Given the potential value of auxiliary variables for bias reduction and efficiency gains, we believe the latter option is preferable in most settings.

Clearly, results based on a restricted simulation study such as this cannot be generalized to all applied settings. For example, in this study we did not consider scenarios with missingness in auxiliary variables, multiple auxiliary variables, or more complex regression models, all of which are common in practice. Further, in all simulation scenarios the association between the auxiliary variable and the probability of missing data in the outcome was fixed; previous research has shown that the strength of this association is an important determinant of the bias associated with failing to include an auxiliary variable in the imputation model (28). While the simulation study illustrates the potential for introducing bias using MID, the extent of this bias will depend on specific characteristics of the individual study. Associations involving auxiliary variables may be weaker than those considered in this study, and hence the bias introduced by MID may not be of practical importance in many settings (22, 28, 29). Alternatively, researchers may have access to a large number of auxiliary variables, which collectively could have a dramatic influence on bias and efficiency. Thus, while the bias involved in estimating regression coefficients with 20% missing data was moderate in the current study, it could be larger in other settings with similar amounts of missing data.

A further limitation of this study is that we only considered a MAR mechanism and a correctly specified imputation model. Both conditions may not be met in practice. Although a MAR assumption is often plausible, data may instead be missing not at random, which occurs when the probability of missingness depends on unobserved values (4). Unless missingness occurs by design, it is impossible to tell whether data are truly MAR or missing not at random based only on observed values. If imputation is performed under a MAR assumption when data are in fact missing not at random, in general this will lead to biased inference, although auxiliary variables can help to mitigate this bias (22). Since MID is unable to incorporate information about an incomplete outcome from auxiliary variables, it may be that this approach would produce more biased estimates than standard MI when data are missing not at random, although this remains to be investigated. In choosing between standard MI and MID, another important consideration is the ability to adequately specify the imputation model. One argument for using MID is that removing imputed outcomes from the analysis will reduce the bias introduced by a misspecified model for imputing outcomes. Whether this is important in practice is unclear. Popular methods of imputation such as multivariate normal imputation and fully conditional specification are known to be fairly robust to model misspecification (20, 23, 30, 31), while ad hoc approaches such as predictive mean matching can be used when there is uncertainty surrounding relationships between variables in the imputation model (23, 32). Thus, even in settings where there is considerable uncertainty in specifying an appropriate imputation model, we would still recommend proceeding with standard MI when auxiliary information for an incomplete outcome is available.

In summary, MID can lead to biased estimation when auxiliary variables that are associated with missingness in an incomplete outcome are included in the imputation model. Once a decision has been made to include auxiliary variables in the imputation model, whether to satisfy a MAR assumption or to improve precision, we recommend retaining this information in the analysis and using a standard MI approach.
ACKNOWLEDGMENTS

Author affiliations: School of Public Health, University of Adelaide, Adelaide, South Australia, Australia (Thomas R. Sullivan, Amy B. Salter, Philip Ryan); Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Royal Children’s Hospital, Melbourne, Victoria, Australia (Katherine J. Lee); and Department of Paediatrics, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Victoria, Australia (Katherine J. Lee).

The work was supported by funding from an Australian Postgraduate Award (T.R.S.) and National Health and Medical Research Council Career Development Fellowship 1053609 (K.J.L.).

Conflict of interest: none declared.

REFERENCES