Received: 6 April 2022 | Revised: 18 October 2022 | Accepted: 3 February 2023

DOI: 10.1002/sim.9685

RESEARCH ARTICLE

Statistics in Medicine WILEY

Impute-then-exclude versus exclude-then-impute: Lessons when imputing a variable used both in cohort creation and as an independent variable in the analysis model

Peter C. Austin[1,2,3] | **Daniele Giardiello**[4] | **Stef van Buuren**[5,6]

[1]ICES, Toronto, Ontario, Canada

[2]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

[3]Sunnybrook Research Institute, Toronto, Ontario, Canada

[4]Institute for Biomedicine (affiliated with the University of Lübeck), Eurac Research, Bolzano, Italy

[5]University of Utrecht, Utrecht, The Netherlands

[6]Netherlands Organisation for Applied Scientific Research TNO, Leiden, The Netherlands

Correspondence
Peter C. Austin, ICES, G106, 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada.
Email: peter.austin@ices.on.ca

Funding information
Canadian Institutes of Health Research, Grant/Award Number: PJT 166161; Heart and Stroke Foundation of Canada, Grant/Award Number: Mid-Career Investigator Award; Ontario Ministry of Health; Ministry of Long-Term Care

We examined the setting in which a variable that is subject to missingness is used both as an inclusion/exclusion criterion for creating the analytic sample and subsequently as the primary exposure in the analysis model that is of scientific interest. An example is cancer stage, where patients with stage IV cancer are often excluded from the analytic sample, and cancer stage (I to III) is an exposure variable in the analysis model. We considered two analytic strategies. The first strategy, referred to as "exclude-then-impute," excludes subjects for whom the observed value of the target variable is equal to the specified value and then uses multiple imputation to complete the data in the resultant sample. The second strategy, referred to as "impute-then-exclude," first uses multiple imputation to complete the data and then excludes subjects based on the observed or filled-in values in the completed samples. Monte Carlo simulations were used to compare five methods (one based on "exclude-then-impute" and four based on "impute-then-exclude") along with the use of a complete case analysis. We considered both missing completely at random and missing at random missing data mechanisms. We found that an impute-then-exclude strategy using substantive model compatible fully conditional specification tended to have superior performance across 72 different scenarios. We illustrated the application of these methods using empirical data on patients hospitalized with heart failure when heart failure subtype was used for cohort creation (excluding subjects with heart failure with preserved ejection fraction) and was also an exposure in the analysis model.

KEYWORDS
missing data, Monte Carlo simulations, multiple imputation

1 | INTRODUCTION

Missing data is common in clinical and epidemiological research. Missing data occurs when the value of a variable is recorded for some subjects in the sample, but not for all subjects. Failure to correctly account for missing data can lead to

estimates with reduced precision and to biased parameter estimates. Multiple imputation (MI), developed by Rubin, is a popular method for addressing missing data.[1] MI entails the use of an imputation model to generate M ($M > 1$) plausible values for each missing observation, resulting in M complete or filled-in datasets. A separate analysis is conducted in each of the M complete datasets, with the estimated quantities (eg, regression coefficients) pooled across the M complete datasets. Unlike single imputation, MI reflects the uncertainty around the missing values, and allows the analyst to recover some of the lost information. For these reasons, MI is increasingly seen in clinical research as the preferred approach for addressing missing data.

An issue that has not been addressed in the methodological literature is how best to impute a variable that is used as an inclusion/exclusion criterion when constructing the analytic sample and that is also an exposure variable of interest in the analysis model that is of scientific interest. Examples of this occur in the cancer literature. Cancer stage describes the size of a given cancer and the degree to which it has spread. One common cancer staging system classifies a given cancer into one of four stages: stage I, stage II, stage III, and stage IV (there is also a stage 0 that denotes the presence of abnormal cells that have not yet become cancerous).[2] Stage IV denotes a cancer that has spread to distant parts of the body. Many cancer studies exclude patients with stage IV cancers, as these patients have a prognosis that is substantially worse than patients with non-stage IV cancers.[3-7] Thus, cancer stage is often used as an inclusion/exclusion criterion when creating a study sample. In the study sample, cancer stage can then be used as an exposure variable in the subsequent analysis model that is of scientific interest. Thus, the outcome (eg, cause-specific mortality) is regressed on cancer stage (stages I to III) and a set of covariates, allowing the analyst to estimate the independent association of cancer stage with the outcome.

Cancer stage is often subject to missingness, so that cancer stage is not documented for some subjects. The likelihood of missing data on cancer stage can depend on type of cancer, issues of data quality, characteristics of patients with poor life expectancy, and characteristics of patients associated with poor access to quality health services.[8] Research has suggested that the proportion of missing data on cancer stage is higher in the elderly, those with high levels of comorbidity or complex care needs, and those in institutionalized settings.[9-11] Similarly, race, gender, marital status, place of residence, and receipt of surgical treatment have also been associated with missing data on cancer stage.[12] In addition, there is evidence that patients with missing data on cancer stage have survival probabilities that lie between regional and advanced (metastatic) disease.[13,14] For these reasons, one can assume that when cancer stage data are missing, that they are not missing completely at random (MCAR). Prior studies have suggested using MI to address the issue of missing data on cancer stage in population-based cancer registries under the assumption that missing data mechanism is missing at random (MAR).[15,16]

The best strategy has not been determined for imputing missing data when a variable that is subject to missingness is used both for inclusion/exclusion criteria and as an independent exposure variable in the subsequent analysis model. In the cancer literature, some studies have used what we refer to as an "exclude-then-impute" strategy.[17-19] Using this strategy, subjects who have documented evidence of a stage IV cancer are excluded from the sample. The resultant sample consists of those subjects with documented evidence of stage I, II, or III cancer or who have missing data on cancer stage. Within this sample, MI is used to fill in missing values of cancer stage. Since the only observed values of cancer stage in this restricted sample are I, II, or III, subjects who have missing cancer stage are restricted to having one of these 3 values imputed to fill in missing values of cancer stage. An alternative strategy can be referred to as "impute-then-exclude." Using this strategy, MI is applied in the full original sample to fill in missing values of cancer stage (allowing some of the filled in values to be stage IV). Then, in each of the completed samples, stage is used as an inclusion/exclusion criterion to exclude those with stage IV cancer. In each of the resultant restricted samples, the analysis model is fit in which the outcome is regressed on cancer stage and a set of covariates. Despite the lack of methodological studies examining the relative merits of these two strategies, the "exclude-then-impute" strategy has been used in the cancer literature.[17-19] From a theoretical perspective, the "impute-then-exclude" strategy would appear to be the better option, as it allows imputed values of stage to take on any of the initially observed values (stage I to stage IV) (assuming those with stage 0 were initially excluded). However, the "impute-then-exclude" approach could lead to bias in the pooled variance estimation due to incompatibility, in that the sample used for imputation differs from the sample used for the substantive analysis (ie, the analysis sample differs from the imputation sample).[20] Given the lack of methodological studies examining these two approaches, and the potential that the "impute-then-exclude" approach can suffer from incompatibility, it is important that this question be addressed rigorously. We note that these two strategies are intended for use in observational studies, rather than in randomized controlled trials, where there is less likely to be missing data on those variables that are used for inclusion/exclusion criteria.

In a recent study that developed a prognostic model for use in women with breast cancer, women with metastatic cancer (equivalent to stage IV cancer) were excluded and then MI was conducted in the resultant restricted sample.[17] In another study examining breast cancer risk factors and survival by tumor subtype, the authors excluded those with stage IV breast cancer, and then used imputation in the resultant restricted sample to impute stage for those with missing data on stage.[18] Similarly, in a recent study comparing laparoscopic versus open surgery for rectal cancer, the authors excluded subjects with stage IV cancer and then imputed stage for those with missing stage in the resultant sample.[19] Given the lack of methodological studies evaluating the performance of the impute-then-exclude and the exclude-then-impute strategies, it is not known what the consequences were of using the exclude-then-impute approach in these studies.

The objective of the current article was to compare the relative performance of two strategies to imputation, exclude-then-impute versus impute-then-exclude, when a variable that is subject to missingness is used both as an inclusion/exclusion criterion for cohort creation and as an exposure variable of interest in the analysis model. The article is structured as follows: In Section 2, we describe a series of Monte Carlo simulations that were used to address this question. In Section 3, we report the results of these simulations. In Section 4, we provide a case study illustrating the application of these two strategies. Finally, in Section 5, we summarize our findings and place them in the context of the existing literature.

2 | MONTE CARLO SIMULATION METHODS

We conducted an extensive series of Monte Carlo simulations to compare the relative performance of "exclude-then-impute" and "impute-then-exclude" when a variable that is subject to missingness is used both for cohort creation and as the primary exposure variable in the analysis model. To simplify the simulations and their reporting, we focus on a setting in which the target variable is a 3-level categorical variable and subjects from one level of this variable are excluded from the analytic sample. We consider three strategies for dealing with missing data: (i) a complete case analysis; (ii) "exclude-then-impute," (iii) "impute-then-exclude." With the latter strategy we consider four different methods for imputing missing data. To simplify the design of the simulations, we explicitly assumed that only this categorical variable is subject to missingness and that the other variables are not subject to missingness.

2.1 | Factors in the Monte Carlo simulations

Let Z denote the 3-level categorical target variable that is used for both cohort creation and as the primary exposure variable in the analysis model. Z takes on three values: 1, 2, and 3, and subjects with $Z = 3$ are excluded from the analysis sample. We allowed two factors to vary in our simulations: (i) $p_{Z=3} = \Pr(Z = 3)$, the proportion of subjects for whom $Z = 3$; (ii) p_{missing}, the proportion of subjects for whom Z is missing. The former took six values: from 0.1 to 0.6 in increments of 0.1, while the latter took 12 values: from 0.05 to 0.60 in increments of 0.05. We thus considered 72 different scenarios (6×12) in a full factorial design.

2.2 | Data-generating process

We simulated data for a super-population of size 1 000 000. For each subject we generated a 3-level exposure variable from a multinomial distribution, such that the probability of each of the first two levels was $\Pr(Z = 1) = \Pr(Z = 2) = 0.5(1 - p_{Z=3})$, while the probability of the third level was $\Pr(Z = 3) = p_{Z=3}$. We then generated a continuous baseline covariate, X, such that the mean of X varied across the levels of Z. X followed a normal distribution with SD equal to 1 within each of the three levels of Z. The mean of X was −0.2, 0, and 0.2 in those with $Z = 1$, $Z = 2$, and $Z = 3$, respectively. In the simulations that follow, the categorical variable Z will be used both as an inclusion/exclusion criterion when creating the analytic sample and as the exposure of interest in the analysis model that is fit in the restricted sample that excludes those with $Z = 3$.

Having generated a categorical exposure variable and a continuous baseline covariate, we then generated a binary outcome Y using a logistic model. This is done in such a way that the log-odds of the outcome varies across the levels of Z and so that the relationship between X and the log-odds of the outcome is different in those with $Z = 3$ than in those with $Z = 1$ or $Z = 2$:

$$\text{logit}(\Pr(Y=1)) = \begin{cases} \beta_0 + \beta_3 X & \text{if } Z = 1, \\ \beta_0 + \beta_1 + \beta_3 X & \text{if Z} = 2, \\ \beta_0 + \beta_2 + (\beta_3 + \beta_4) X & \text{if Z} = 3. \end{cases} \quad (1)$$

We have incorporated an interaction between the categorical variable Z and the continuous variable X, such that the slope for X differs between those with $Z = 3$ and those with $Z = 1$ and $Z = 2$. The inclusion of this interaction was motivated by the cancer stage variable in the cancer literature. As noted above, subjects with stage IV cancers are often excluded from studies since their prognosis is very different from those subjects with stage I, II, or III cancers. Furthermore, the association of covariates with the outcome may differ in those with stage IV cancers from the association in those with other stages. The true values of the regression coefficients were: $\beta_0 = 0.25$, $\beta_1 = \log(0.75)$, $\beta_2 = \log(0.50)$, $\beta_3 = \log(1.5)$, and $\beta_4 = \log(2) - \log(1.5)$. In the restricted sample consisting of those subjects for whom $Z = 1$ or $Z = 2$, the analysis model of interest is $\text{logit}(\Pr(Y = 1)) = \beta_0 + \beta_1 I(Z = 2) + \beta_3 X$. Thus, there is a common slope for X, while the odds of the outcome differs for those with $Z = 2$ compared to those with $Z = 1$, after adjusting for X.

Having generated Z (a 3-level categorical variable), X (a continuous covariate), and Y (a binary outcome), we then induced missing data in the super-population (we also retained a copy of the super-population in which no data were missing). We induced missingness only in Z. This was done using two different mechanisms: (i) MCAR; (ii) MAR.[1] Under each missing data mechanism, the proportion of subjects with missing data was equal to p_{missing}. Under a MCAR mechanism, the probability that Z was set to missing was equal to p_{missing} for all subjects regardless of their values of X, Z, and Y. Under a MAR mechanism, the likelihood that Z was set to missing was related to both X and Y (but not to the value of Z itself, otherwise the data would be missing not at random [MNAR]). In the missing data model, the weight for X was twice the weight for Y.

2.3 | Analyzes in the simulated datasets

We drew a random sample of size 1000 from the super-population. This was done twice, first from the super-population with no missing data (as noted above a copy of the super-population was retained in which no data were missing), second from the super-population that was subject to missing data.

In the random sample with no missing data (drawn from the super-population in which no data were set to missing), we excluded subjects with $Z = 3$ and fit the analysis model, in which the binary outcome was regressed on both Z and X using a logistic regression model. The estimate of the three regression coefficients (the intercept, and the coefficients for Z and X) and their standard errors were extracted. We refer to this analysis as the "no missing data" analysis. This reflects the analysis that would be conducted had missing data not occurred. This analysis will serve as the "gold standard" against which the other strategies will be compared.

We now describe analyzes for the three strategies for dealing with missing data: a complete case analysis, "exclude-then-impute" and "impute-then exclude." The first strategy was to conduct a complete case analysis. In the random sample that was subject to missingness we excluded subjects with missing data. In this restricted sample, we then excluded subjects with $Z = 3$. In this further restricted sample, we fit the analysis model as above. We refer to this strategy as the "complete case" strategy.

The second strategy was "exclude-then-impute." In the random sample that was subject to missingness we excluded those subjects for whom $Z = 3$. In the resultant restricted sample, MI was used to fill in missing values of Z. Both X and Y were included in the imputation model. We created M completed versions of the sample, where M was set equal to the percentage of subjects in the sample for whom Z was missing.[21] In each complete dataset the analysis model described above was fit. Regression coefficients and their standard errors were pooled across the M imputed datasets using Rubin's Rules.[1] We refer to this strategy as "exclude-then-impute."

The third strategy was "impute-then-exclude." In the random sample that was subject to missingness we used MI to fill in missing values for Z. We created M completed versions of the sample, where M was equal to the percentage of subjects in the sample for whom Z was missing. In each completed dataset we then excluded those subjects for whom $Z = 3$ and fit the analysis model described above in the resultant restricted sample. Note that a given subject for whom Z was missing, could have an imputed value of $Z = 3$ in some of the imputed datasets and $Z \neq 3$ in other imputed datasets. Consequently, this subject will be excluded from some of the analysis samples and included in some of the analysis samples. Thus there may be variation in the size of the complete datasets. Regression coefficients and their standard

errors were pooled across the M imputed datasets using Rubin's Rules. Within the "impute-then-exclude" strategy we considered four different methods to impute missing values of Z. First, only X and Y were included as main effects in the imputation model. We refer to this method as "impute-then-exclude (no interaction)." Second, a multiplicative interaction between X and Z was created in the random sample (if Z were missing then the value of the interaction term would be missing as well). In this second method of imputation, X, Y, and the interaction between X and Z were included in the imputation model for Z. The interaction between X and Z was also imputed using the "just another variable" (JAV) approach to imputing interactions.[21] This method is referred to as "impute-then-exclude" (interaction - JAV). The third method to imputation was similar to the second, except that passive imputation was done for the missing interaction terms between X and Z (ie, when Z was missing, its value was imputed and then the product X^*Z was computed). This method is referred to as "impute-then-exclude" (interaction - passive). The fourth method used substantive model compatible fully conditional specification (SMCFCS) to impute missing values.[22] This method requires the specification of an analysis model of interest. Since we are performing MI in the full sample prior to exclusion of those with $Z = 3$, we used the analysis model that was used to generate outcomes in the full sample (this model is represented by formula (1) above). Note that for this method, the analysis model in the full sample differs from the model of scientific interest, which is fit only in those with $Z = 1$ or $Z = 2$. We refer to this method as "impute-then-exclude (interaction - SMCFCS)."

We thus considered three strategies, the latter of which included four different methods: complete case analysis, "exclude-then-impute," "impute-then-exclude (no interaction)," "impute-then-exclude (interaction-JAV)," "impute-then-exclude (interaction - passive)," and "impute-then-exclude (interaction - SMCFCS)." Using each of these six strategies/methods, we obtained the estimated regression coefficients and their associated standard errors (pooled across the M complete datasets using Rubin's Rules). Using these two quantities, 95% confidence intervals for the regression parameters were constructed in each of the simulated datasets using standard normal-theory methods. We repeated the above process 1000 times, so that our simulations used 1000 simulation replicates. For each of the 6 methods, we determined the mean of each of the three regression coefficients in the analysis model (β_1, β_2, and β_3) across the 1000 simulation replicates. The relative bias in each estimated coefficient was determined by: $100 \times \frac{\frac{1}{1{,}000}\left(\sum_{i=1}^{1{,}000} \widehat{\beta}_j^i - \beta_j\right)}{\beta_j}$, $j = 0, 1, 3$, where $\widehat{\beta}_j^i$ denotes the pooled estimate of the jth regression coefficient in the ith simulation replicate. We also estimated the empirical coverage rate of estimated 95% confidences as the proportion of constructed confidence intervals that contained the true value of the parameter.

The above process was done once using a MCAR missing data mechanism and once using a MAR missing data mechanism.

2.4 | Software

The simulations were conducted using the R statistical programming language (version 3.6.3).[23] The missing data mechanism was implemented using the ampute function from the mice package (version 3.13.0),[24] while multiple imputation using the MICE algorithm was done using the mice function. SMCFCS was implemented using the SMCFCS function from the SMCFCS package (version 1.4.0).[22]

3 | MONTE CARLO SIMULATION RESULTS

We summarize our results separately for the two missing data mechanisms (MCAR vs. MAR).

3.1 | MCAR missing data mechanism

The relative bias in estimating the regression coefficients is reported in Figure 1 (β_0 - the regression intercept), Figure 2 (β_1 - the regression slope for $Z = 2$), and Figure 3 (β_3 - the regression slope for X). Each figure consisted of six panels, one for each of the values of $\Pr(Z = 3)$. On each panel we have superimposed a horizontal line denoting a relative bias of 0%.

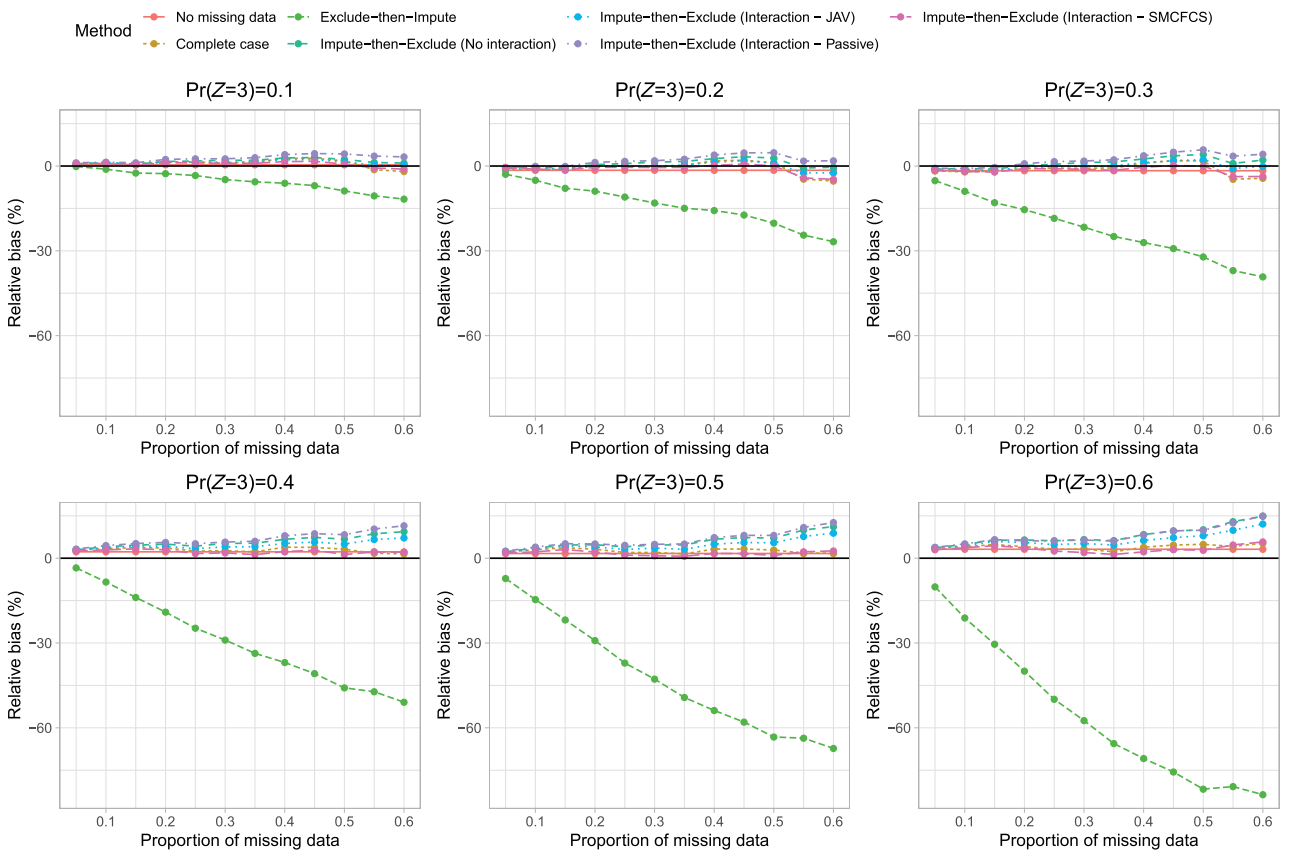

FIGURE 1 Relative bias in estimating β_0 (intercept) (MCAR)

The exclude-then-impute strategy tended to result in the greatest relative bias in estimating the intercept (β_0), with the relative bias increasing with prevalence of missing data and as $\Pr(Z = 3)$ increased. Impute-then-exclude (no interactions; interactions - JAV; interactions - passive) resulted in modest bias when the prevalence of missing data was high and $\Pr(Z = 3) \geq 0.40$. The complete case analysis and impute-then-exclude (interaction - SMCFCS) tended to result in estimates of the intercept with minimal bias across all scenarios. The increasing bias observed for the exclude-then-impute strategy may be explained by noting that the intercept in a logistic regression model is related to the probability of the outcome for those with $X = 0$ in the reference category of Z. Using an exclude-then-impute strategy, subjects for whom Z is missing have imputed values equal to either $Z = 1$ or $Z = 2$, whereas some of these subjects truly had $Z = 3$. The value of $\beta_2 = \log(0.50)$ (the regression coefficient for $Z = 3$) in the data-generating process implies that subjects with $Z = 3$ have a higher probability of the outcome than do subjects in the reference level ($Z = 1$). Incorrectly including subjects who truly have $Z = 3$, but whose imputed value of Z is $Z = 1$, in the analysis sample will result in the estimated intercept being biased. As the proportion of subjects with missing data increases, an increasing proportion of subjects will be misclassified.

When estimating the slope for $Z = 2$ (β_1), most methods tended to result in estimates with modest bias when both the prevalence of missing data was >0.50 and $\Pr(Z = 3) \geq 0.40$.

When estimating the slope for X (β_3), the complete case analysis and impute-then-exclude (interaction - SMCFCS) tended to result in estimates with minimal bias. With the other four methods, bias increased with both the increasing prevalence of missing data and as $\Pr(Z = 3)$ increased. A potential explanation for the bias observed when using exclude-then-impute is similar to that provided above. Of those subjects who truly have $Z = 3$, for whom there is a different X slope, those with missing data will be imputed as having $Z = 1$ or $Z = 2$. The inclusion of these subjects will therefore result in bias in estimating the common slope for $Z = 1$ and $Z = 2$. Three of four impute-then-exclude methods (excluding the interaction - SMCFCS strategy) likely resulted in bias in the estimated slope for X because the imputation model was not able to account for the fact that in the original full sample, the slope for X differed in those with $Z = 3$ than it did in those with $Z = 1$ or $Z = 2$. Thanks to the specification of the correct complete-data analysis model, only the SMCFCS strategy was able to account for this, thereby not introducing any subsequent bias.

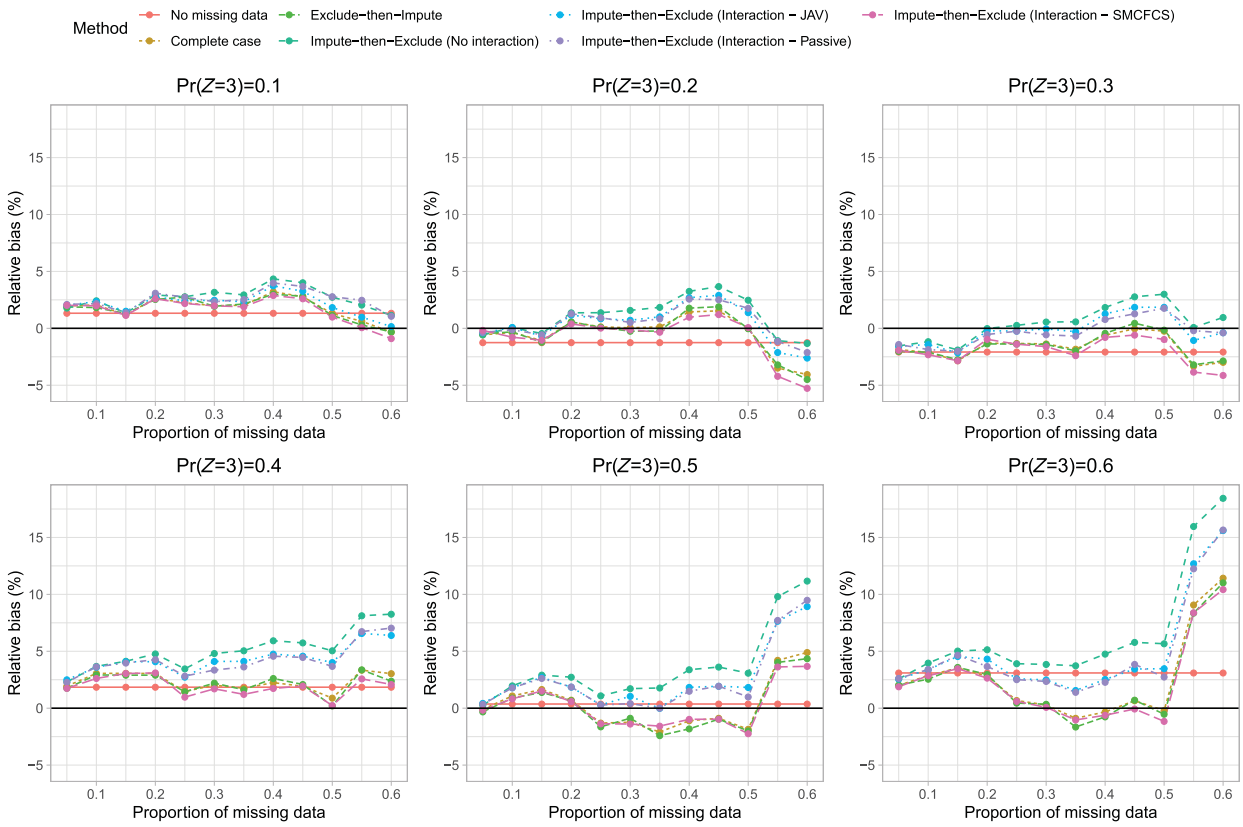

FIGURE 2 Relative bias in estimating β_1 (slope for $Z = 2$) (MCAR)

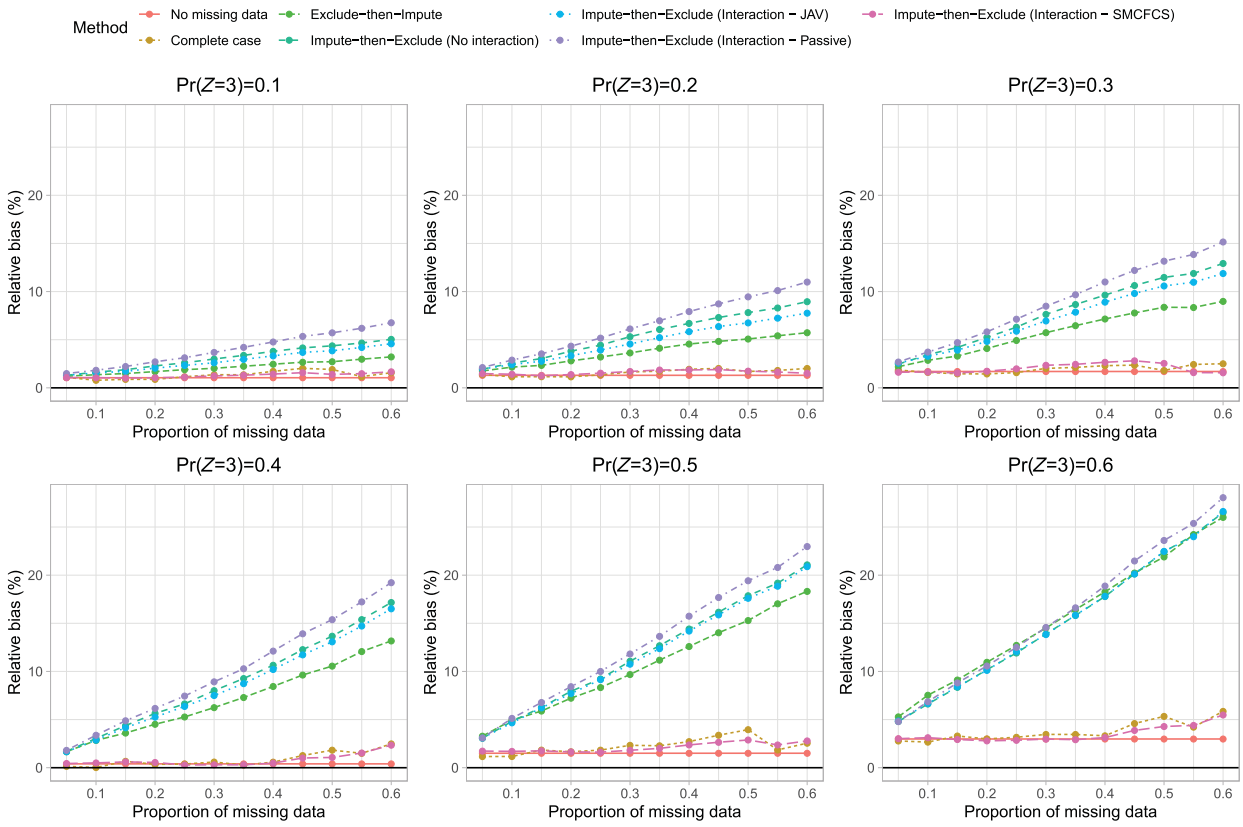

FIGURE 3 Relative bias in estimating β_3 (slope for X) (MCAR)

Empirical coverage rates of estimated 95% confidence intervals are reported in Figure 4 (confidence intervals for β_1, the slope for $Z = 2$) and Figure 5 (confidence intervals for β_3, the slope for X). Each figure consists of six panels, one for each value of $\Pr(Z = 3)$. On each panel we have superimposed a horizontal line denoting the advertised coverage rate of 95%. Due to our use of 1000 simulation iterations, an empirical coverage rate that is less than 93.65% or greater than 96.35% would be statistically significantly different from the advertised rate based on a standard normal-theory test. We have added two additional horizontal lines denoting these two thresholds.

Confidence intervals for β_1 tended to have the correct coverage rate except when the prevalence of missing data was high and one of three methods were used: impute-then-exclude (no interaction), impute-then-exclude (interaction - JAV), and impute-then-exclude (interaction - passive). In this case, empirical coverage rates tended to be below the advertised rate. We hypothesize that these lower than advertised coverage rates were due to the minor bias in estimating β_1 that was observed above (Figure 2).

Confidence intervals for β_3 tended to have the advertised coverage rate except with the exclude-then-impute strategy when $\Pr(Z = 3) \geq 0.40$. Impute-then-exclude (interaction - passive) had coverage rates slightly below the advertised rate when $\Pr(Z = 3) \geq 0.40$ and the prevalence of missing data was very high. In exclude-then-impute the proportion of imputation errors (ie, imputing $Z = 2$ instead of true $Z = 3$) rises with $\mathrm{P}(Z = 3)$. Thus, in imputed $Z = 2$ we see more Y events than there should be. This contributes to the bias in the intercept β_0, and potentially, as a side effect, increases the probability of events to a range where the variance of the binomial becomes too small, leading to confidence intervals that are too short.

In examining Figures 1-3, we note that there is a minor relative bias for the "no missing data" analysis in many of the panels. However, the magnitude of the bias varies across panels. Furthermore, in Figures 4 and 5, the empirical coverage rate for this analysis is never statistically significantly different from the advertised rate of 95%. This suggests that the observed relative bias for the "no missing data" analysis in any given scenario simply reflects random variation. We hypothesize that, with larger sample sizes, the magnitude of the relative bias would decrease across scenarios.

3.2 | MAR missing data mechanism

Relative bias in estimating the regression coefficients is reported in Figure 6 (β_0 - the regression intercept), Figure 7 (β_1 - the regression slope for $Z = 2$), and Figure 8 (β_3 - the regression slope for X). These figures have a similar structure to Figures 1 to 3.

When estimating the regression intercept, we do not show the relative bias for the complete case analysis, as its relative bias was as large as 200%. Including it on the figures made it difficult to differentiate between the performance of the other strategies. Of the different strategies, impute-then-exclude (interaction - SMCFCS) tended to have the best performance, followed by impute-then-exclude (interaction - passive). When using impute-then-exclude (no interaction) or impute-then-exclude (interaction - JAV), relative bias tended to increase as $\Pr(Z)$ increased and as the prevalence of missing data increased. Exclude-then-impute tended to perform poorly when $\Pr(Z = 3) \geq 0.4$. The large bias in estimating β_0 when using the complete case analysis is likely due to the fact that under a MAR missing data mechanism, the complete case sample is not a representative sample of included subjects (ie, of those with $Z = 1$ or $Z = 2$) and that the prevalence of the outcome differs in the complete cases compared to the prevalence of the outcome in the full analysis sample if there were no missing data. Subjects with missing data who truly had $Z = 3$ will have a higher prevalence of the outcome. However, these subjects will have imputed values of Z that are either $Z = 1$ or $Z = 2$, resulting in biased estimates of the regression intercept.

When estimating the slope for $Z = 2$, impute-then-exclude (interaction - JAV) tended to result in the largest bias, with the relative bias increasing as $\Pr(Z = 3)$ increased and as the prevalence of missing data increased. The remaining methods tended to have comparable performance to one another. We are unsure as to why the JAV approach performs poorly compared to the passive imputation approach. We would note that not all statisticians would prefer the JAV approach over the passive imputation approach.

When estimating the slope for X, the complete case strategy resulted in the largest relative biases, with a relative bias that tended to increase with increasing prevalence of missing data. This bias is likely because, under a MAR missing data mechanism, the complete case analysis sample is a nonrepresentative sample of the full analysis sample. Subjects with missing data who truly had $Z = 3$ will have a different slope for X than will subjects with $Z = 1$ or $Z = 2$. However, these subjects will have imputed values of Z that are either $Z = 1$ or $Z = 2$, resulting in biased estimates of the regression slope for X. Of the remaining methods, impute-then-exclude (interaction - SMCFCS) tended to result in essentially unbiased

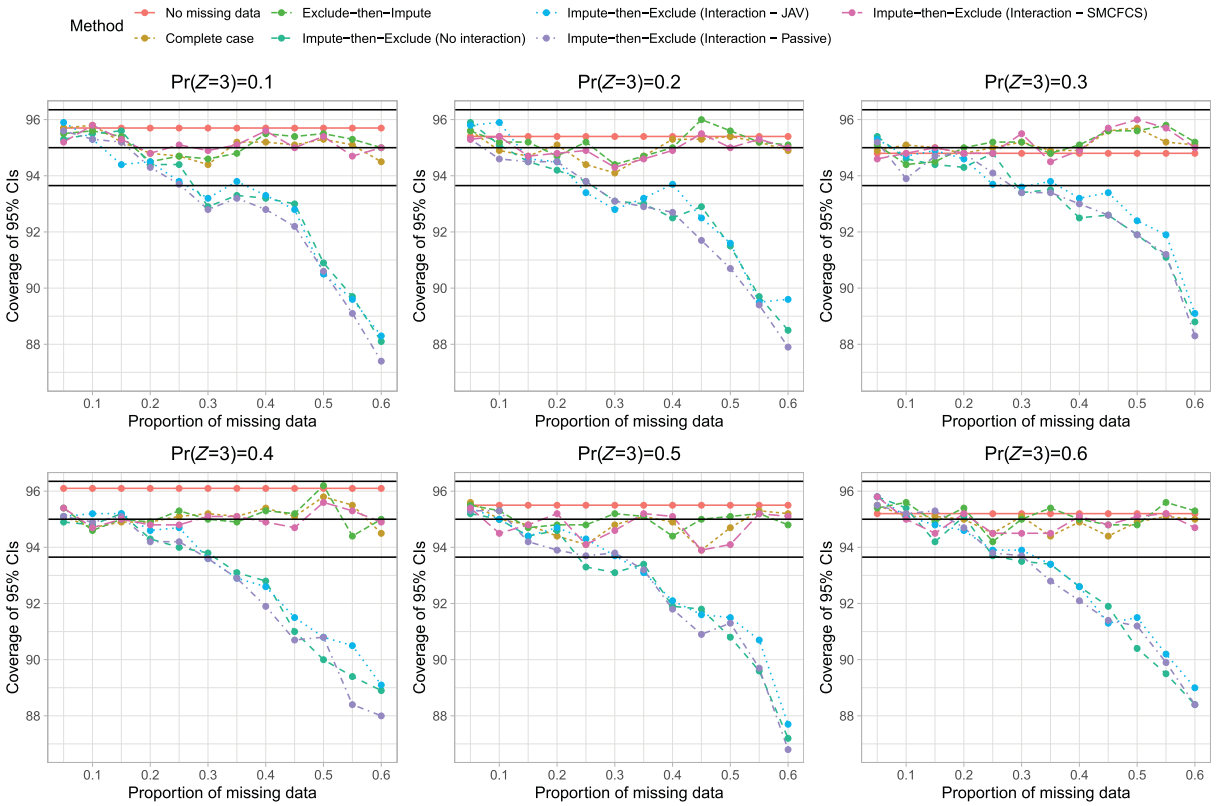

FIGURE 4 Coverage of 95% CIs for β_1 (slope for $Z = 2$) (MCAR)

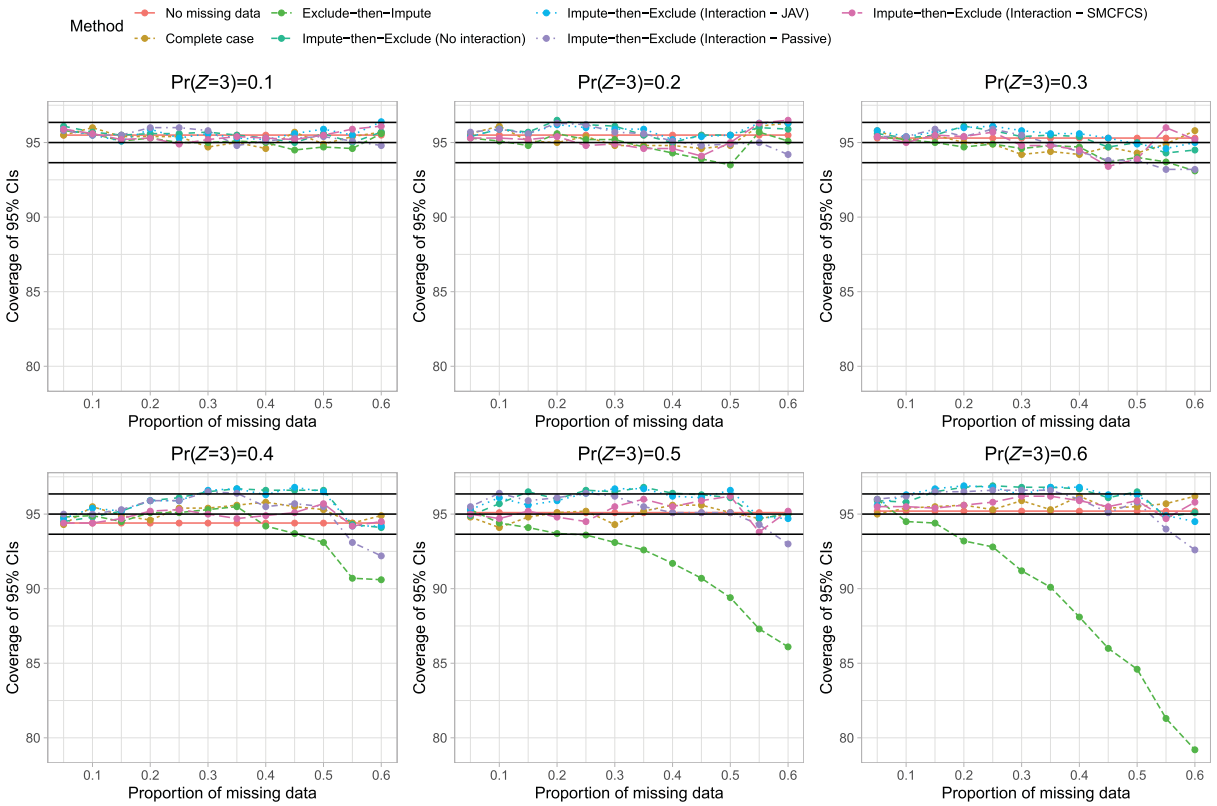

FIGURE 5 Coverage of 95% CIs for β_3 (slope for X) (MCAR)

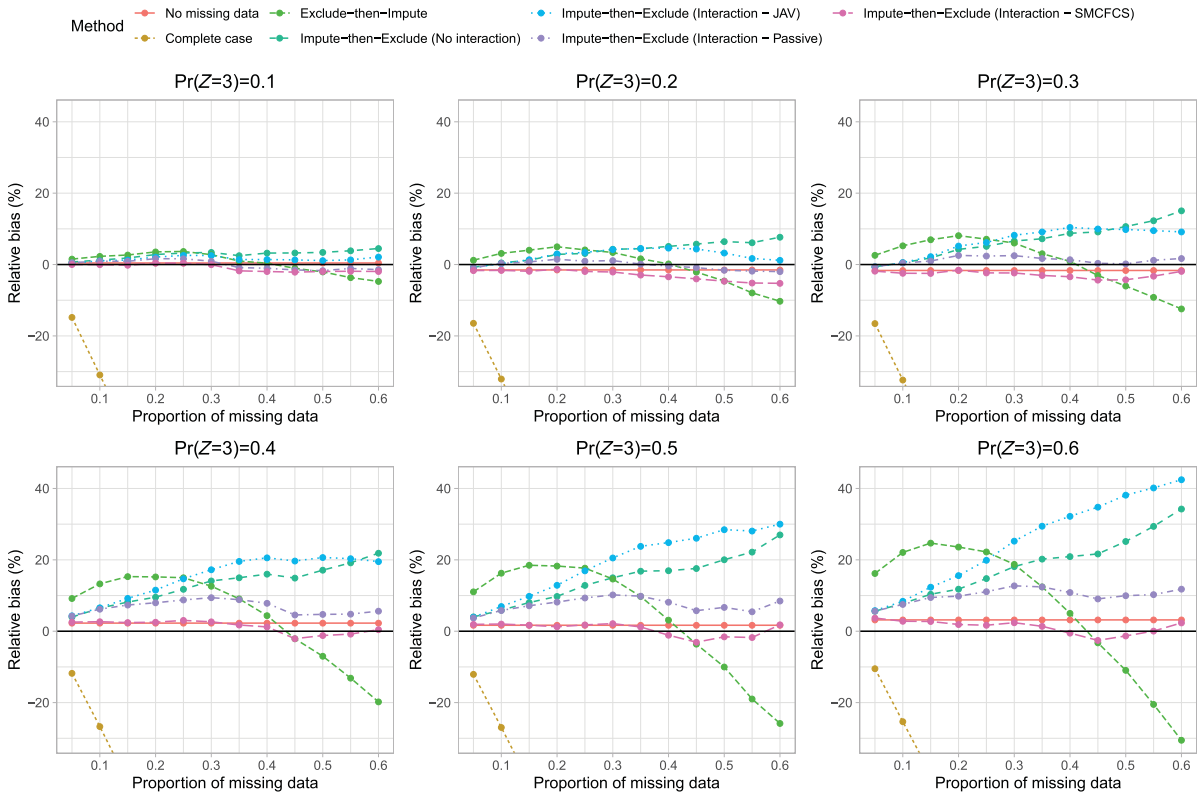

FIGURE 6 Relative bias in estimating β_0 (intercept) (MAR)

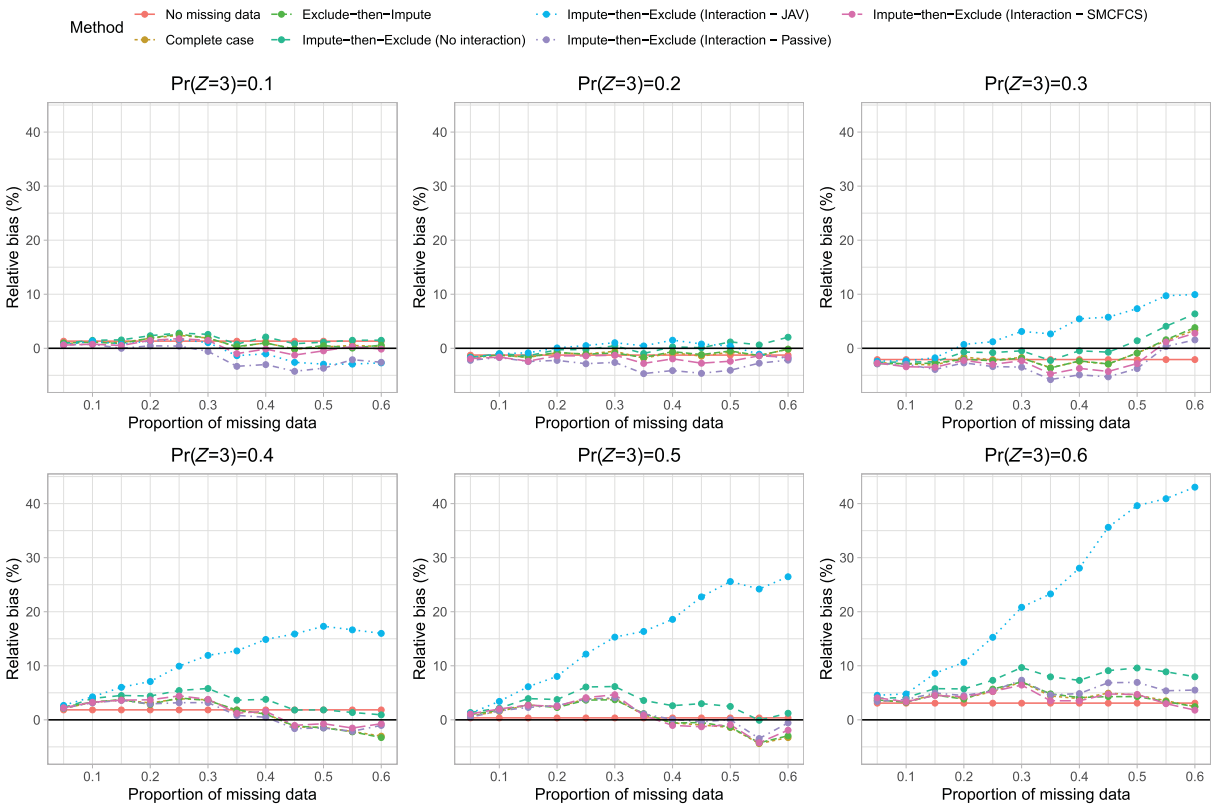

FIGURE 7 Relative bias in estimating β_1 (slope for $Z = 2$) (MAR)

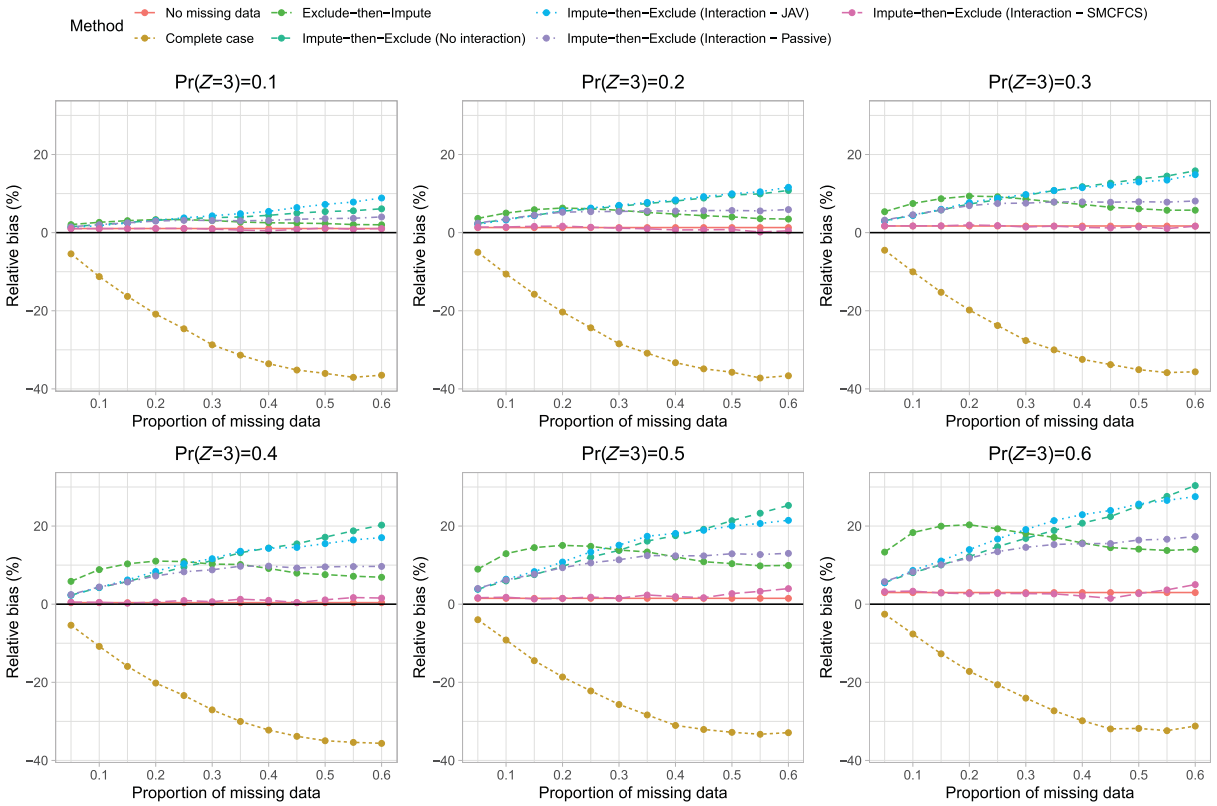

FIGURE 8 Relative bias in estimating β_3 (slope for X) (MAR)

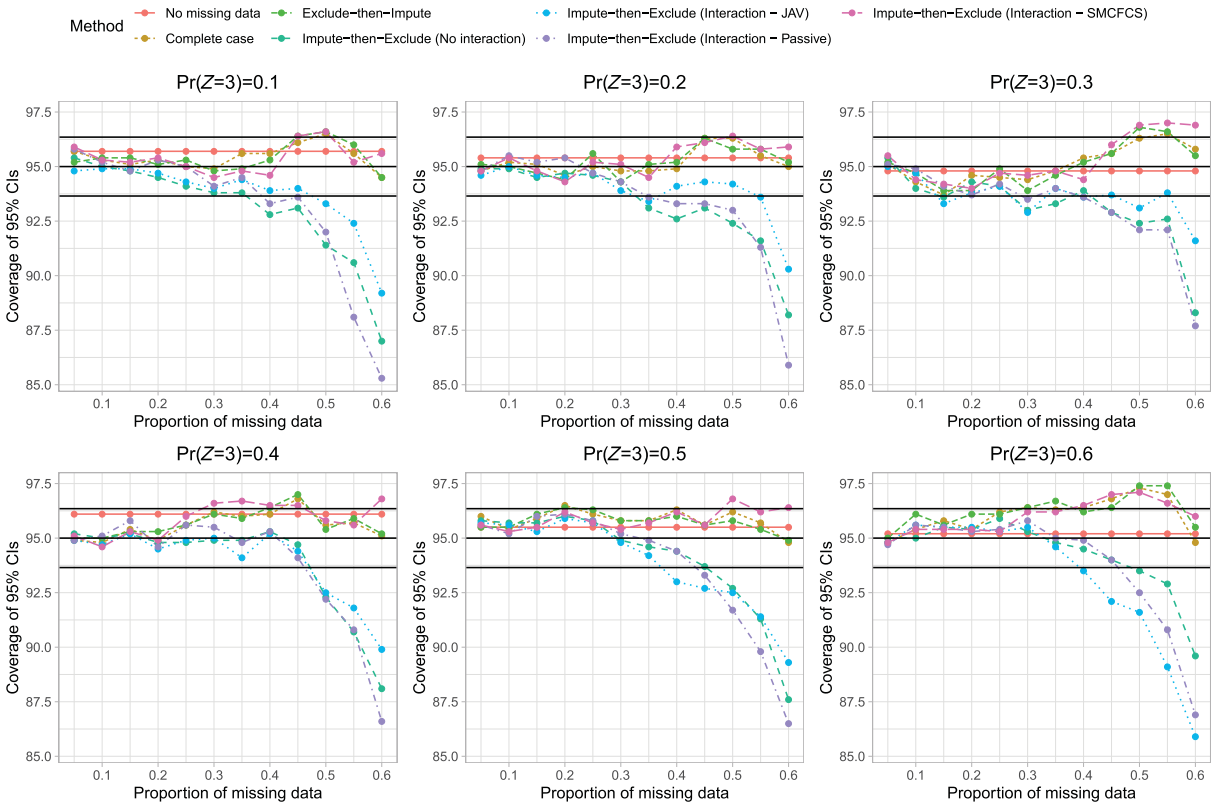

FIGURE 9 Coverage of 95% CIs for β_1 (slope for $Z = 2$) (MAR)

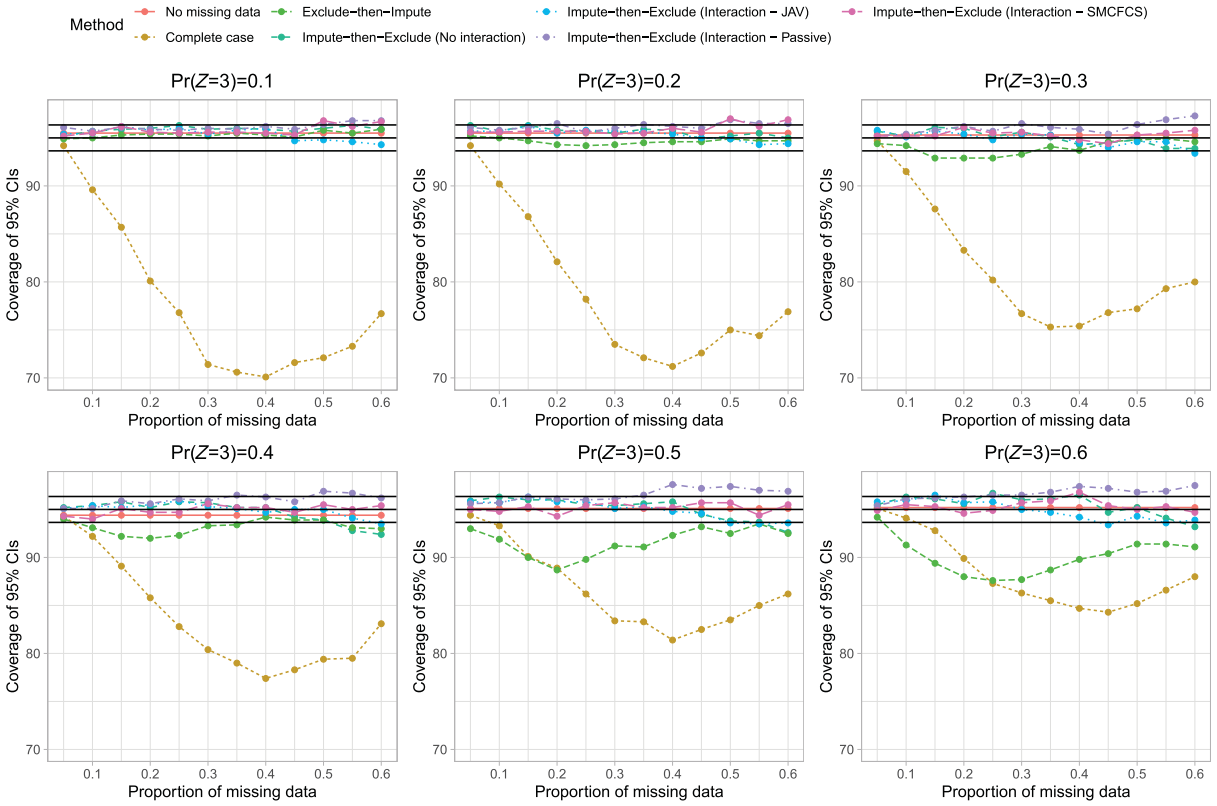

FIGURE 10 Coverage of 95% CIs for β_3 (slope for X) (MAR)

estimates. Exclude-then-impute tended to result in bias that increased as $\Pr(Z = 3)$ increased. Impute-then-exclude (no interaction) and impute-then-exclude (interaction - JAV) resulted in increasing bias as both $\Pr(Z = 3)$ and the prevalence of missing data increased. A similar phenomenon was observed for impute-then-exclude (interaction - passive), although it tended to result in less bias than the former two methods.

Empirical coverage rates of estimated 95% confidence intervals are reported in Figure 9 (confidence intervals for β_1, the slope for $Z = 2$) and Figure 10 (confidence intervals for β_3, the slope for X). These figures have a similar structure to Figures 4 and 5.

Confidence intervals for β_1 tended to have the correct coverage rate except when the prevalence of missing data was high and one of three strategies were used: impute-then-exclude (no interaction), impute-then-exclude (interaction - JAV), and impute-then-exclude (interaction - passive). In this case, empirical coverage rates tended to be below the advertised rate.

Confidence intervals for β_3 tended to have the advertised coverage rate except when either of two strategies was used: complete case analysis or exclude-then-impute (with the latter tending to have the advertised coverage rate when $\Pr(Z = 3) \leq 0.3$). The suboptimal coverage rates when using the complete case analysis are likely due to the substantial bias in estimating β_3 observed above (Figure 8).

4 | CASE STUDY

We provide a case study to illustrate the application of the strategies described above. The case study consisted of patients hospitalized with heart failure (HF). HF is classified into one of three types: HF with reduced ejection fraction (HFrEF), HF with a mid-range ejection fraction (HFmEF), and HF with preserved ejection fraction (HFpEF). Ejection fraction (EF) is typically measured using echocardiography, which is not done for all patients, and so HF type is subject to missingness. In some instances, EF is not measured because the hospital to which the patient was admitted had no facilities for echocardiography. In other instances, EF is not assessed due to chance alone. Rarely, the patient died before echocardiography could be done. Given that the imputation model included death and age (which is associated with

death), it is reasonable to assume that missing data are MAR. Neither age nor death are subject to missingness as they can be obtained from linkage with provincial registries that contain these data for all residents of Ontario.

Our objective is to form a restricted sample consisting of those with HFrEF and HFmEF (ie, excluding those with HFpEF) and then use logistic regression to regress death within 1 year on HF type (HFrEF vs. HFmEF) and age.

4.1 | Methods

The initial sample consisted of 9943 patients hospitalized with a diagnosis of HF. These data were from the first phase of the enhanced feedback for effective cardiac treatment (EFFECT) study,[25] which collected data on patients hospitalized with HF in Ontario, Canada between 1999 and 2001. Of these patients, 2544 (26%) had documented HFrEF, 633 (6%) had documented HFmEF, 1261 (13%) had documented HFpEF, while HF type was missing for 5505 (55%).

The outcome was a binary variable denoting death within 1 year of hospital admission. Of the 9943 patients, 3297 (33.2%) died within 1 year of hospital admission. The analysis model was a logistic regression model in which the binary outcome was regressed on HF type and age.

We used the six strategies described above to create an analysis sample consisting of those with HFrEF and HFmEF. When using MI, we created M complete samples, where M was the percentage of subjects with missing data. When using exclude-then-impute, we used $M = 63$, while when using impute-then-exclude, we used $M = 55$. For a given strategy (except the complete case strategy), a separate logistic regression model was fit in each of the M complete samples and the estimated regression coefficients were pooled using Rubin's Rules.

4.2 | Results

The estimated regression coefficients for the intercept, the log-odds ratio for HFmEF versus HFrEF, and the log-odds ratio for age (per 10-year increase in age) are reported in Figure 11. There is one panel each of the three regression coefficients.

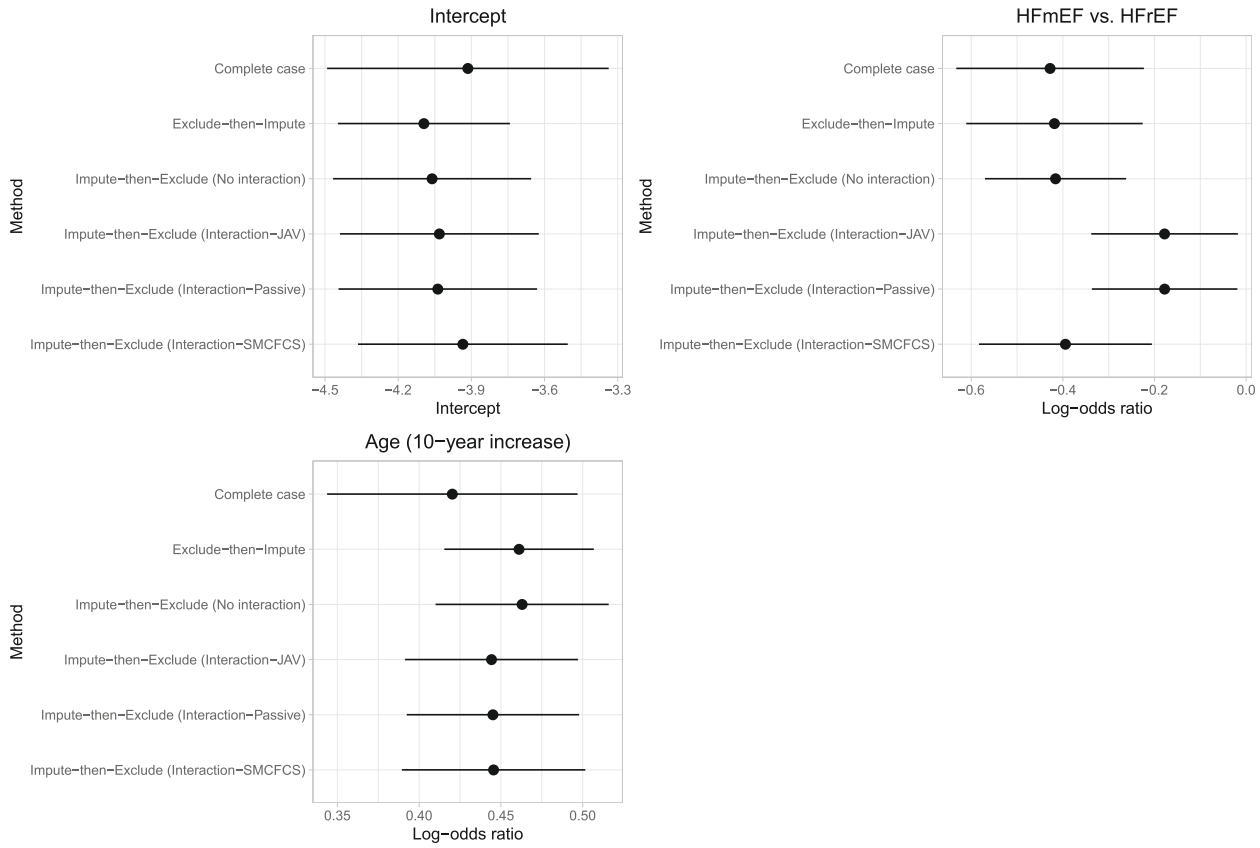

FIGURE 11 Coefficient estimates from case study

Each panel consists of a forest plot depicting the estimated regression coefficient and its associated 95% confidence interval obtained using each of the six strategies.

The six strategies resulted in qualitatively similar estimates of the intercept. The point estimate ranged from −4.09 (exclude-then-impute) to −3.91 (complete case analysis).

When estimating the log-odds ratio for HFmEF versus HFrEF, the use of impute-then-exclude (Interaction - JAV) and impute-then-exclude (interaction - passive) resulted in estimates that were very similar to one another (approximately −0.2, equivalent to an odds ratio of 0.82), whereas the other four methods resulted in estimated log-odds ratios that were close to −0.4, which is equivalent to an odds ratio of approximately 0.67.

When estimating the log-odds ratio for a 10-year increase in age, the complete case analysis resulted in a log-odds ratio of 0.42, corresponding to an odds ratio of 1.52. The other five strategies produced log-odds ratios of approximately 0.45, corresponding to odds ratios of approximately 1.57.

The greatest observable difference between strategies was when estimating the effect of HFmEF versus HFrEF. Interestingly, impute-then-exclude (interaction with JAV or passive) produced qualitatively different estimates than the other four methods. Based on the simulations, we would suggest that the estimate produced by impute-then-exclude (interaction - SMCFCS) most likely produced the estimate that was the closest to the truth.

5 | DISCUSSION

We examined the setting in which a variable that is subject to missingness is used both as an inclusion/exclusion criterion in creating the analytic cohort and as an exposure variable in the analysis model that is of scientific interest. We compared the relative performance of a strategy based on exclude-then-impute with that of four methods based on a strategy of impute-then-exclude. These strategies were also compared with a complete case analysis strategy. We found that a strategy based on impute-then-exclude using SMCFCS imputation in the original sample tended to have very good performance across a wide range of scenarios.

We are unaware of previous methodological studies that have examined how to account for missing data in variables that are used both for inclusion/exclusion criteria and that are included as exposures in the subsequent analysis model. As discussed in the Introduction, such variables are common in the cancer literature, where cancer stage is often used to exclude subjects with stage IV cancer from the analytic sample. Our case study included data from cardiology, where heart failure subtype is an important variable that is both subject to missingness and can that be used as an inclusion/exclusion criterion when creating the analytic sample. In the case study, data on heart failure subtype was missing for 55% of subjects. A study by Giganti and Shepherd examined the setting in which a variable that is subject to missingness is used for a study exclusion criterion.[20] In their simulations, they considered a setting with a continuous variable that was subject to missingness, and that subjects whose values of this continuous variable were greater than a specified threshold were included in the analysis sample (whereas we considered a setting with a categorical variable that was subject to missingness and was used for study inclusion/exclusion). They compared the Rubin's Rules estimate of the variance of the pooled estimate with that of an estimator proposed by Robins and Wang.[26] They found that the Rubin's Rules estimator resulted in estimated 95% confidence intervals that were conservative. The variance estimator proposed by Robins and Wang is "consistent even when the imputation and analysis models are misspecified and incompatible with one another."[26] We did not consider the Robins and Wang variance estimator for two reasons. First, we did not consider this estimator due to the lack of existing statistical software packages for obtaining this estimate. As noted by Hughes and colleagues, this estimator "has the potential to provide more robust inferences, should the considerable challenges in provision of software implementing the procedure be overcome."[27] Similarly, Giganti and Shepherd state that "creating software that generalized the RW estimator would be a challenging but worthy endeavor."[20] Given the lack of software for implementing this method, we did not pursue it in the current study. Second, the Robins and Wang estimator requires that the imputation model is a fully parametric probability model.[26] Such a full parametric probability model is not, in general, provided by the mice algorithm that we used for imputing missing data. Our use of Rubin's Rules for estimating the variance of the estimated regression coefficients is unlikely to be too problematic in this setting. While Giganti and Shepherd found that the Robins and Wang estimator was more efficient (ie, the estimated variance was smaller) than the Rubin's Rules estimator, efficiency is unlikely to be the cause of the biases that we observed. Furthermore, Giganti and Shepherd found that the use of Rubin's Rules resulted in estimated 95% confidence intervals that were conservative (ie, the empirical coverage rates exceeded the advertised rate). We would judge this to be a lesser problem than having confidence intervals whose empirical coverage rates were lower than the advertised rate. We suggest that the use of Rubin's Rules was satisfactory

in the current study as the estimated confidence achieved at least the advertised coverage rates and a smaller variance estimate would have been unlikely to affect our conclusions.

The current study is subject to certain limitations. First, we relied on the use of Monte Carlo simulations to examine the relative performance of different strategies. Due to the computational complexity of our simulations, we were restricted to considering a limited number of scenarios (eg, the simulations under the MAR scenarios when using impute-then-exclude (interaction - passive) required 237 h (9.9 days) while the simulations using the impute-then-exclude (interaction - JAV) method under the MAR scenarios required 310 h (12.9 days)). However, our 72 scenarios included a wide range of prevalences of missing data (from 0.05 to 0.60 in increments of 0.05) and a wide range of prevalences of level of the categorical variable that was used for excluding subjects from the analytic sample (from 0.10 to 0.60 in increments of 0.10). These 72 scenarios were sufficient to illustrate that there were scenarios in which some strategies resulted in substantial bias. For instance, the use of impute-then-exclude (interaction - JAV) resulted in biased estimation of both the log-odds ratio for the categorical variable and the continuous variable when $\Pr(Z = 3)$ was at least 0.40 and the prevalence of missing data was moderate to high under a MAR missing data mechanism. Similarly, there were scenarios in which the exclude-then-impute strategy led to biased estimation of the log-odds ratio for the continuous variable. Because of these computational limitations, we limited our focus to settings in which the categorical variable that was subject to missingness was a 3-level categorical variable. We also restricted our focus to settings in which the sample size was 1000. It is possible that the performance of SMCFCS may deteriorate in settings in which the categorical variable has more than 3 levels or when the sample size is different. Second, in both the simulations and the case study, we assumed that both the outcome and the other baseline covariates were not subject to missingness. In the case study this was a realistic assumption, as these variables were obtained from provincial registries that contain demographic data on all residents of Ontario. In different settings, other covariates may be subject to missingness. We hypothesize that, if the data are MAR and an appropriate imputation model is used, then the results of the current simulations would hold. Third, we restricted our examination to two different missing data mechanisms: MCAR and MAR. We did not examine settings in which the missing data mechanism for the 3-level categorical variable was missing not at random (MNAR). We hypothesize that, were the missing data mechanism to be MNAR, the different methods that were examined would fail to perform adequately. Consistent with many studies examining the performance of MI, we have restricted our focus to MCAR and MAR missing data mechanisms.

While we commented on several specific observations when summarizing the results of the Monte Carlo simulations above, there are a few overarching themes that merit further discussion. First, estimation of β_0 (the intercept) and β_3 (the slope for X) were biased with the complete case analysis under the MAR mechanism for missing data, but not under the MCAR mechanism for missing data. These differences in bias are due entirely to the missing data mechanism. Under a MCAR analysis, those with complete data are representative of the entire sample. Thus, a complete case analysis would not be expected to result in bias. Second, in many scenarios bias increased with increasing prevalence of missing data and as the $\Pr(Z = 3)$ increased. Third, we frequently observed that bias increased and coverage decreased as the proportion of missing data increased and as $\Pr(Z = 3)$ increased. For three of the impute-then-exclude methods (all but the interactions - SMCFCS approach), we suggest that this is due to incompatibility between the imputation model and the analysis model. For the exclude-then-impute strategy, we suggest that this is due to systematic misclassification in Z in subjects for whom Z was missing. Of those subjects who had missing Z, some truly had $Z = 3$. However, for all of these subjects, the imputed value of Z would be either 1 or 2 when using exclude-then-impute. The inclusion of these subjects in the analysis sample would then introduce bias in the estimation of the coefficients of the analysis model (which would also affect coverage rates of 95% confidence intervals). The degree of systematic misclassification would increase as the proportion of missing data increased and as $\Pr(Z = 3)$ increased. Fourth, impute-then-exclude (SMCFCS) was observed to perform well across all scenarios. We suggest that the good performance of this method was due to the specification of an outcomes model in the full sample that was consistent with the analysis model in the restricted sample. By including this additional information, we were able to resolve the issue of incompatibility between the imputation sample and the analysis sample.

The success of the impute-then-exclude strategy with the SMCFCS method for imputation depends on the correct specification of the outcome (or analysis) model in the full sample. In both simulation and application, we allowed for different slopes across the three groups. In the simulation context we know that the assumption is correct, but in the application it remains an assumption. If the outcome model would have (incorrectly) specified equal slopes across the three levels of Z (ie, a main effects model), then the results from SMCFCS would have been no better than the "impute-then-exclude" no-interaction method. In the analytic problem of this article, the outcome model in the full sample differs from the model of scientific interest, which includes only $Z = 1$ and $Z = 2$. Thus, in practice the information on the nature of

the relations for subjects with $Z = 3$ could be scant as these subjects are generally removed before analysis. Outside the simulation context, it may not always be obvious how to specify the outcome model in the full sample. It may help to perform a preliminary analysis on the data using only $Z = 3$ subjects, which may suggest the need for an interaction (ie, regressing the binary outcome on X in those subjects with $Z = 3$ can provide an indication of whether the slope for X is the same in these subjects as in those with $Z = 1$ or $Z = 2$). Assuming as little as possible about the data seems like a wise strategy in general, but—when taken to the extreme—also will weaken the impact of the analysis (or outcome) model in the full sample.

In summary, we recommend that researchers consider the use of a strategy of impute-then-exclude using SMCFCS with a carefully specified full-data model when faced with a variable that is subject to missing data that is used both for inclusion/exclusion criteria and as an exposure variable in the analysis model.

ACKNOWLEDGEMENTS

ICES is an independent, non-profit research institute funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). As a prescribed entity under Ontario's privacy legislation, ICES is authorized to collect and use health care data for the purposes of health system analysis, evaluation and decision support. Secure access to these data is governed by policies and procedures that are approved by the Information and Privacy Commissioner of Ontario. The opinions, results and conclusions reported in this article are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOH or MLTC is intended or should be inferred. The use of data in this project was authorized under section 45 of Ontario's Personal Health Information Protection Act, which does not require review by a research ethics board. This research was supported by operating grant from the Canadian Institutes of Health Research (CIHR) (PJT 166161).

DATA AVAILABILITY STATEMENT

The dataset from this study is held securely in coded form at ICES. While legal data sharing agreements between ICES and data providers (e.g., healthcare organizations and government) prohibit ICES from making the dataset publicly available, access may be granted to those who meet pre-specified criteria for confidential access, available at http://www.ices.on.ca/DAS (email: das@ices.on.ca).

ORCID

Peter C. Austin https://orcid.org/0000-0003-3337-233X
Daniele Giardiello https://orcid.org/0000-0002-9005-9430

REFERENCES

1. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons; 1987.
2. Accessed February 14, 2022. https://www.cancer.gov/about-cancer/diagnosis-staging/staging.
3. Giardiello D, Kramer I, Hooning MJ, et al. Contralateral breast cancer risk in patients with ductal carcinoma in situ and invasive breast cancer. *NPJ Breast Cancer*. 2020;6(1):60.
4. Akdeniz D, Kramer I, van Deurzen CHM, et al. Risk of metachronous contralateral breast cancer in patients with primary invasive lobular breast cancer: results from a nationwide cohort. *Cancer Med*. 2022. doi:10.1002/cam4.5235
5. Kramer I, Schaapveld M, Oldenburg HSA, et al. The influence of adjuvant systemic regimens on contralateral breast cancer risk and receptor subtype. *J Natl Cancer Inst*. 2019;111(7):709-718.
6. Kramer I, Hooning MJ, Mavaddat N, et al. Breast cancer polygenic risk score and contralateral breast cancer risk. *Am J Hum Genet*. 2020;107(5):837-848.
7. Akdeniz D, van Barele M, Heemskerk-Gerritsen BAM, et al. Effects of chemotherapy on contralateral breast cancer risk in BRCA1 and BRCA2 mutation carriers: a nationwide cohort study. *Breast*. 2022;61:98-107.
8. Gurney J, Sarfati D, Stanley J, et al. Unstaged cancer in a population-based registry: prevalence, predictors and patient prognosis. *Cancer Epidemiol*. 2013;37(4):498-504.
9. di Girolamo C, Walters S, Benitez Majano S, et al. Characteristics of patients with missing information on stage: a population-based study of patients diagnosed with colon, lung or breast cancer in England in 2013. *BMC Cancer*. 2018;18(1):492.
10. Worthington JL, Koroukian SM, Cooper GS. Examining the characteristics of unstaged colon and rectal cancer cases. *Cancer Detect Prev*. 2008;32(3):251-258.
11. Ramos M, Franch P, Zaforteza M, Artero J, Duran M. Completeness of T, N, M and stage grouping for all cancers in the Mallorca cancer registry. *BMC Cancer*. 2015;15:847.
12. Merrill RM, Sloan A, Anderson AE, Ryker K. Unstaged cancer in the United States: a population-based study. *BMC Cancer*. 2011;11:402.

13. Nur U, Shack LG, Rachet B, Carpenter JR, Coleman MP. Modelling relative survival in the presence of incomplete data: a tutorial. *Int J Epidemiol*. 2010;39(1):118-128.
14. Nederland IK. Dutch; 2022. Accessed October 4, 2022. Available at: https://iknl.nl/nkr-cijfers?fs%7Cepidemiologie_id=527&fs%7Ctumor_id=1&fs%7Coverlevingssoort_id=531&fs%7Cklassificatie_stadium_id=639&fs%7Cstadium_id=685%2C684%2C683%2C682%2C687&fs%7Cjaren_na_diagnose_id=688%2C689%2C690%2C691%2C692%2C693%2C694%2C695%2C696%2C697%2C698%2C699&cs%7Ctype=line&cs%7CxAxis=jaren_na_diagnose_id&cs%7Cseries=stadium_id&ts%7CrowDimensions=stadium_id&ts%7CcolumnDimensions=jaren_na_diagnose_id&lang%7Clanguage=nl
15. Eisemann N, Waldmann A, Katalinic A. Imputation of missing values of tumour stage in population-based cancer registration. *BMC Med Res Methodol*. 2011;11:129.
16. Luo Q, Egger S, Yu XQ, Smith DP, O'Connell DL. Validity of using multiple imputation for "unknown" stage at diagnosis in population-based cancer registry data. *PLoS One*. 2017;12(6):e0180033.
17. Giardiello D, Steyerberg EW, Hauptmann M, et al. Prediction and clinical utility of a contralateral breast cancer risk model. *Breast Cancer Res*. 2019;21(1):144.
18. Morra A, Jung AY, Behrens S, et al. Breast cancer risk factors and survival by tumor subtype: pooled analyses from the breast cancer association consortium. *Cancer Epidemiol Biomarkers Prev*. 2021;30(4):623-642.
19. Dehlaghi Jadid K, Cao Y, Petersson J, Angenete E, Matthiessen P. Long term oncological outcomes for laparoscopic versus open surgery for rectal cancer - a population-based nationwide noninferiority study. *Colorectal Dis*. 2022;24:1308-1317.
20. Giganti MJ, Shepherd BE. Multiple-imputation variance estimation in studies with missing or misclassified inclusion criteria. *Am J Epidemiol*. 2020;189(12):1628-1632.
21. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377-399.
22. Bartlett JW, Seaman SR, White IR, Carpenter JR, Alzheimer's Disease Neuroimaging Initiative. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat Methods Med Res*. 2015;24(4):462-487.
23. *R: a Language and Environment for Statistical Computing [Computer Program]*. Vienna: R Foundation for Statistical Computing; 2005.
24. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1-67.
25. Tu JV, Donovan LR, Lee DS, et al. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *JAMA*. 2009;302(21):2330-2337.
26. Robins JM, Wang N. Inference for imputation estimators. *Biometrika*. 2000;87:113-124.
27. Hughes RA, Sterne J, Tilling K. Comparison of imputation variance estimators. *Stat Methods Med Res*. 2016;25(6):2541-2557.

How to cite this article: Austin PC, Giardiello D, van Buuren S. Impute-then-exclude versus exclude-then-impute: Lessons when imputing a variable used both in cohort creation and as an independent variable in the analysis model. *Statistics in Medicine*. 2023;1-17. doi: 10.1002/sim.9685