Handling missing values in the analysis of between-hospital differences in ordinal and dichotomous outcomes

Handling missing values in the analysis of between-hospital differences in ordinal and dichotomous outcomes: a simulation study

Reinier C A van Linschoten ,[1,2,3] Marzyeh Amini ,[1] Nikki van Leeuwen,[1] Frank Eijkenaar,[4] Sanne J den Hartog ,[1,5,6] Paul J Nederkoorn,[7] Jeannette Hofmeijer,[8,9] Bart J Emmer,[10] Alida A Postma,[11,12] Wim van Zwam,[11] Bob Roozenbeek,[5] Diederik Dippel,[5] Hester F Lingsma,[1] MR CLEAN Registry Investigators

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2023-016387).

For numbered affiliations see end of article.

Correspondence to
Dr Reinier C A van Linschoten, Public Health, Erasmus MC, Rotterdam 3015, Netherlands; r.vanlinschoten@erasmusmc.nl

Received 30 May 2023
Accepted 30 August 2023
Published Online First 21 September 2023

To cite: van Linschoten RCA, Amini M, van Leeuwen N, *et al*. *BMJ Qual Saf* 2023;**32**:742–749.

ABSTRACT

Missing data are frequently encountered in registries that are used to compare performance across hospitals. The most appropriate method for handling missing data when analysing differences in outcomes between hospitals with a generalised linear mixed model is unclear. We aimed to compare methods for handling missing data when comparing hospitals on ordinal and dichotomous outcomes. We performed a simulation study using data from the Multicentre Randomised Controlled Trial of Endovascular Treatment for Acute Ischaemic Stroke in the Netherlands (MR CLEAN) Registry, a prospective cohort study in 17 hospitals performing endovascular therapy for ischaemic stroke in the Netherlands. The investigated methods for handling missing data, both case-mix adjustment variables and outcomes, were complete case analysis, single imputation, multiple imputation, single imputation with deletion of imputed outcomes and multiple imputation with deletion of imputed outcomes. Data were generated as missing completely at random (MCAR), missing at random and missing not at random (MNAR) in three scenarios: (1) 10% missing data in case-mix and outcome; (2) 40% missing data in case-mix and outcome; and (3) 40% missing data in case-mix and outcome with varying degree of missing data among hospitals. Bias and reliability of the methods were compared on the mean squared error (MSE, a summary measure combining bias and reliability) relative to the hospital effect estimates from the complete reference data set. For both the ordinal outcome (ie, the modified Rankin Scale) and a common dichotomised version thereof, all methods of handling missing data were biased, likely due to shrinkage of the random effects. The MSE of all methods was on average lowest under MCAR and with fewer missing data, and highest with more missing data and under MNAR. The 'multiple imputation, then deletion' method had the lowest MSE for both outcomes under all simulated patterns of missing data. Thus, when estimating hospital effects on ordinal and dichotomous outcomes in the presence of missing data, the least biased and most reliable method to handle these missing data is 'multiple imputation, then deletion'.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Missing values are common in quality registries, but no recommendation exists on methods to handle missing data when comparing hospitals on outcomes.

WHAT THIS STUDY ADDS

⇒ All studied methods for handling missing data (complete case analysis; single imputation; single imputation, then deletion; multiple imputation; and multiple imputation, then deletion) lead to biased hospital estimates when comparing hospitals in terms of outcomes using generalised linear mixed models due to shrinkage of the random effects.

⇒ The least biased and most reliable method for handling missing data when estimating hospital effects is the 'multiple imputation, then deletion' method.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ For future quality initiatives that compare hospitals on outcomes with generalised linear mixed models, the 'multiple imputation, then deletion' method is recommended to handle missing data.

⇒ Caution remains warranted when the percentage of missing data increases or when a missing not at random mechanism is suspected.

BACKGROUND

Benchmarking of hospital performance is a promising tool to improve healthcare, but a first requirement is comprehensive, accurate and complete data.[1 2] Although a lot of effort is being put into collecting complete data, even well-designed registries suffer from missing data.[3 4] Missing data can occur in the outcomes of interest, but also in patient characteristics that are used for adjustment, so-called case-mix variables. When measuring hospital performance by estimating hospital effects on outcome in registry data, missing data can introduce bias and reduce precision in several ways.[5–7] First, hospital effects that are estimated from smaller samples will have larger variance. Consequently, hospitals with a large proportion of missing data are less likely to be identified as significantly better or worse than the average. Second, patients with complete information tend to be systematically different from patients with missing data,[8] which can reduce validity of the hospital effect estimates. Third, the reason why data are missing may be related to specific characteristics of the hospitals being compared, for example, with respect to data collection arrangements.[9] Consequently, missing data might confound between-hospital comparisons on outcome.

Generally speaking, three missing data mechanisms can be distinguished: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Under MCAR, values are missing in a random subset of patients. When missing data are related to other observed patient data, missing data are MAR. For example, data would be MAR when women are more likely to have missing data than men. When missing data are related to unobserved data such as the value of the missing data itself, data are MNAR.[10 11] An example of data being MNAR would be patients in poorer health not responding to a quality of life questionnaire.

Several methods for handling missing data have been described.[12–15] Complete case analysis (CCA), single imputation (SI) and multiple imputation (MI) are techniques that are frequently used. These methods are computationally straightforward, versatile, relatively easy to apply and available in standard statistical software programs.[12] CCA is a deletion-based method, in which observations with any missing value are excluded from the analysis. With SI, a regression model with the variable of interest (ie, the one with missing value(s)) as dependent variable and patient, treatment and hospital characteristics as independent variables is fitted. Missing values are then imputed using the value as predicted by the model. With SI, the uncertainty around the predictions generated by the regression is ignored in the further analysis; the imputed data are analysed as they were observed. MI is similar to SI in that a regression model is used to model missing data. This method differs from SI in that it generates more than one data set with different, yet plausible, values for the missing observations. These data sets can separately be analysed and the results pooled. In this way, MI does account for uncertainty around the predictions generated by the regression model for missing values.[15] An extension of SI and MI is a method called 'imputation, then deletion'. With this method, data are first imputed by SI or MI. Then, the imputed outcomes are deleted from the imputed data sets and patients who initially had missing outcome data are not included in the analysis. Imputed values in the case-mix variables are kept. This can improve efficiency compared with MI, and is more robust to misspecification of the imputation regression model.[16]

Despite missing data often being a problem in between-hospital comparisons, no recommendation exists on which of the aforementioned methods to handle missing data to use when estimating hospital effects. In this simulation study, we aimed to assess the bias and reliability of these methods in estimating between-hospital differences in outcome.

METHODS

Design

Study population

Simulations were based on the observational data from the Multicentre Randomised Controlled Trial of Endovascular Treatment for Acute Ischaemic Stroke in the Netherlands (MR CLEAN) Registry, of which the methodology has been published.[17] In brief, this large nationwide stroke registry is a prospective observational cohort study in all 17 hospitals that perform endovascular therapy (EVT) for acute ischaemic stroke in the Netherlands. For the registry extensive clinical and neuroimaging data are collected. We aimed to compare hospital effects on outcome based on different methods of handling missing data with those from a complete reference data set. Therefore, we selected patients from the registry with complete data for the variables in the model described below.

Variables

Our simulation involved the analysis of an ordinal and dichotomous outcome. The ordinal outcome variable was the modified Rankin Scale (mRS).[18] The mRS is a commonly used measure of patients' functional outcome after ischaemic stroke and ranges from 0 (no symptom) to 6 (death). The dichotomous outcome variable was good functional outcome, defined as an mRS of 0–2.[19] The mRS was assessed at 90 days after EVT (±14 days). For case-mix adjustment, we used the following baseline variables: age, sex and the National Institutes of Health Stroke Scale (NIHSS). The NIHSS measures stroke-related neurological deficits and ranges from 0 (no stroke symptom) to 42 (severe stroke). To improve model stability, age and NIHSS were standardised. We also used two process measures as auxiliary variables in the imputation procedure. Auxiliary variables are variables

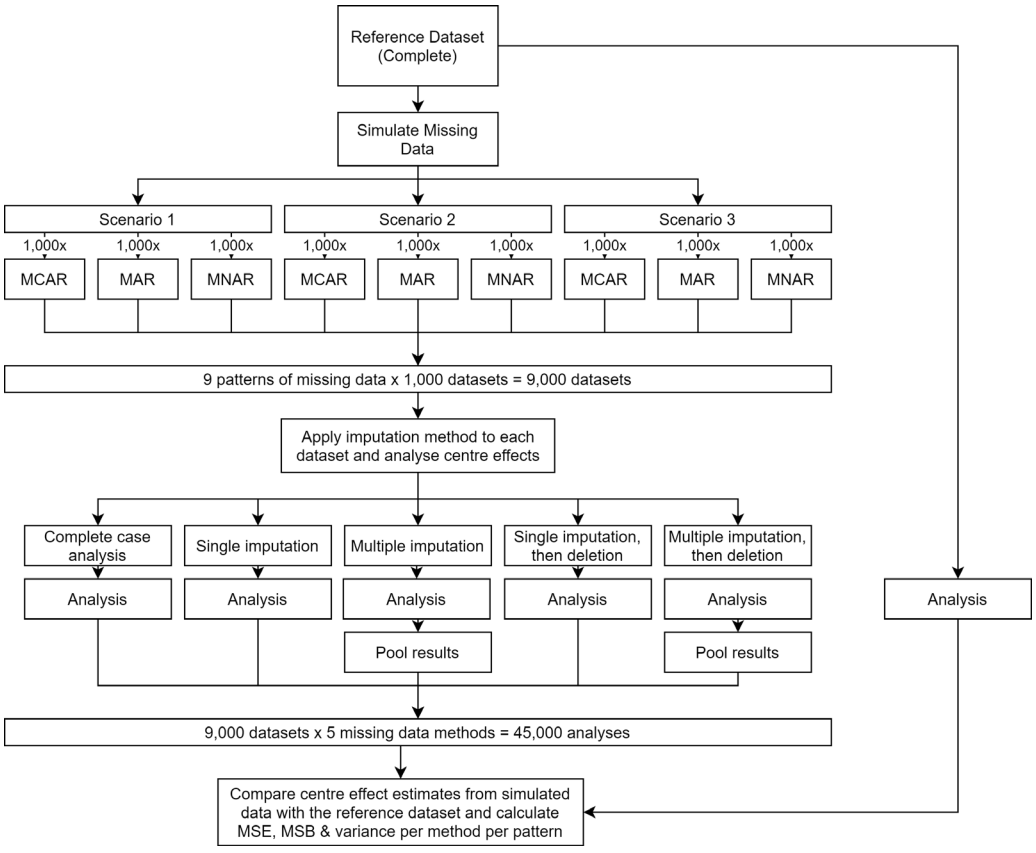

Figure 1 Flow chart of the study methods. Scenario 1: 10% of data missing in both case-mix (age, sex and baseline National Institutes of Health Stroke Scale (NIHSS)) and outcome (modified Rankin Scale, mRS) in all 17 hospitals. Scenario 2: 40% of data missing in the outcome and baseline NIHSS equally distributed over all hospitals. Scenario 3: 40% of data missing in the outcome and baseline NIHSS with the probability of missing data varying between hospitals. MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; MSB, mean squared bias; MSE, mean squared error.

that contain information about the incomplete variables and can improve imputation.[16 20] These variables were time from onset to groin puncture and the use of general anaesthesia during the EVT procedure as these are likely related to the outcome.[21]

Data-generating mechanism

The analysis consisted of three steps (figure 1). In the first step, missing data were simulated assuming three different mechanisms of missing data: MCAR, MAR and MNAR. Multivariate missing data were simulated using the mice package in R.[22] Details of the procedure are described in the online supplemental methods. In addition, we simulated three different scenarios for the degree of missing data. In scenario 1, 10% of data on both case-mix (age, sex and baseline NIHSS) and outcome (ordinal and dichotomous mRS) were made missing in all 17 hospitals. In scenario 2, we assumed 40% of missing data in the outcome and in one important case-mix variable (NIHSS) at baseline, equally distributed over all hospitals. Scenario 3 was the same as scenario 2, but with the probability of data being missing varying between hospitals. The relatively high proportion of 40% missing data per variable in scenarios 2 and 3 was based on findings from Dutch and American quality registries for acute stroke care.[3 4]

Within each scenario, we simulated cases with different combinations of missing data (ie, a patient could have missing data in the outcome, the NIHSS or in both for scenario 2). Missingness of data in one variable was independent of data being missing in another variable. Applying these combinations means that the proportion of patients with missing data was higher than the proportion of missing data per variable. For example, in scenario 2 (with 40% of missing data per variable), 24% of patients had missing data in only the outcome, 24% in only the NIHSS and 16% in both the NIHSS and outcome. The percentage of patients with missing data was 34.39% in scenario 1 and 64% in both scenarios 2 and 3. See also online supplemental tables 1 and 2.

Combining the three mechanisms and three scenarios led to nine patterns of missing data. To account for random variation, each combination was simulated 1000 times.

Estimands

The estimands of our simulation study were the best linear unbiased predictions (BLUPs), that is, the

random effects for the 17 hospitals with their corresponding variance.

Methods for handling missing data

In the analysis step, we applied five methods to handle missing data. These were CCA, SI, MI, 'single imputation, then deletion' (SID) and 'multiple imputation, then deletion' (MID). Imputation was performed with the mice package in R.[22] A binomial generalised linear mixed model was used to impute the dichotomous outcome, while predictive mean matching based on a linear mixed model was used to impute the ordinal outcome. For further details on the imputation procedure, see the online supplemental methods.

Hereafter, generalised linear mixed models were used to estimate between-hospital differences in outcome for each imputation method (figure 1). A model with fixed effects for age, sex and baseline NIHSS, a random effect for hospital and the mRS as the dependent variable was fitted separately for the ordinal and dichotomous versions. For the ordinal outcome, we used a cumulative link mixed model with the mRS reversed, so that a positive fixed or random effect indicates higher odds of a better outcome.[23] For the dichotomous outcome, a binomial generalised linear mixed model was used.[24] Hospital effect estimates were estimated as contrasts between all 17 hospitals using the BLUPs from the generalised linear mixed models. Estimates from the multiple imputed data sets were pooled using Rubin's rules.[15]

Performance measures

The hospital effect estimates from these models were used to assess the bias and reliability of the five methods for handling missing data across the nine patterns of missing data (three scenarios times three mechanisms of missing data). Bias was measured by mean squared bias (MSB), which is the mean difference between the estimates of the hospital effect from the simulated data sets and the hospital effect estimate from the reference data set, squared, and averaged over hospitals.[25] The MSB can be interpreted as the average squared distance between the estimated random effect and the random effect from the reference data set. A higher MSB implies more bias. Reliability was defined as the variance of hospital effect estimates from the simulated data around the mean random effect from the simulated data, averaged over hospitals. This can be interpreted as the spread of random effects, with a higher variance indicating lower reliability. Together, MSB and variance sum to mean squared error (MSE), defined as the mean of the squared differences between hospital effect estimates from the simulated data set and the hospital effect estimates from the reference data set.[25] MSB, variance and MSE were all calculated on the linear predictor scale. When an estimator is

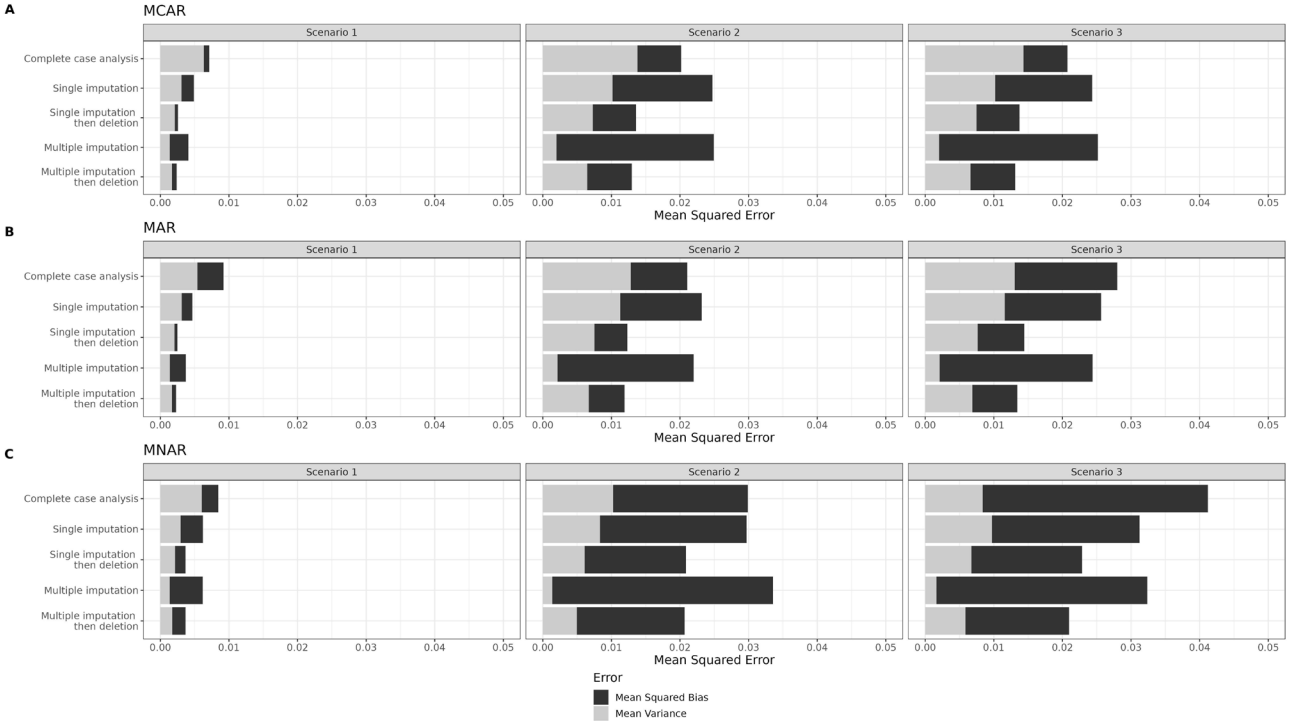

Figure 2 Mean squared error, mean squared bias and mean variance per method for each scenario under (A) MCAR, (B) MAR and (C) MNAR for an ordinal outcome. Scenario 1: 10% of data missing in both case-mix (age, sex and baseline National Institutes of Health Stroke Scale (NIHSS)) and outcome (modified Rankin Scale, mRS) in all 17 hospitals. Scenario 2: 40% of data missing in the outcome and baseline NIHSS equally distributed over all hospitals. Scenario 3: 40% of data missing in the outcome and baseline NIHSS with the probability of missing data varying between hospitals. MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random.

unbiased, the MSB is zero and the MSE equals the variance of the estimator. We also assessed the coverage of the 95% prediction intervals around the hospital effects for the different missing data methods. We use the term prediction interval instead of CI, as the hospital effect estimates are predictions (ie, BLUPs).[26]

To assess uncertainty around the performance measures we calculated the Monte Carlo error of the performance measures. The Monte Carlo error was defined as the SE of the performance measure, estimated using a jackknife resampling approach.[27]

RESULTS

A total of 2817 patients were included in the reference data set (online supplemental table 3). Mean patient age ranged from 66 to 75 years and differed among hospitals (p=0.0018). The percentage of female patients did not vary significantly across hospitals (range 44–67%, p=0.82). Hospitals also differed in mean baseline NIHSS score (13–17, p<0.0001), mean time to groin (183–253, p<0.0001) and use of general anaesthesia (0–99%, p=0.0005). Outcome differed between hospitals; the percentage of patients with a good functional outcome varied from 35% to 51% (p<0.0015). The ORs for the reference hospital effect estimates ranged from 0.71 to 1.65 for the ordinal outcome and from 0.70 to 1.79 for the dichotomous outcome; see online supplemental tables 4 and 5 for the reference effect estimates and 95% prediction intervals.

Exploring the raw results, we found that under all scenarios and with all methods, hospital effects were on average underestimated and there was considerable variance in the estimation of the hospital effects (online supplemental figures 1–10). CCA failed to estimate a hospital effect for one centre in 1 of the 1000 simulations under MNAR for scenario 3 for both the ordinal and dichotomous outcomes. This happened because this hospital had a relatively low sample size (n=18) and all patients had one or more missing data points in these simulations.

The MSE ranged from 0.0023 to 0.041 for the ordinal outcome and from 0.0026 to 0.034 for the dichotomous outcome (see figures 2 and 3 and online supplemental tables 6 and 7 for details). The MSE was on average lowest in scenario 1 and under MCAR, and highest in scenario 3 and under MNAR. Overall, the 'imputation, then deletion' method performed best for both outcomes, in all scenarios and under all mechanisms of missing data. MID slightly outperformed SID, while the standard imputation methods were often outperformed by CCA. In scenario 1 and in scenario 3 under MAR and MNAR the standard imputation methods outperformed CCA for an ordinal outcome. For the dichotomous outcome, the standard imputation methods only outperformed CCA in scenario 1.

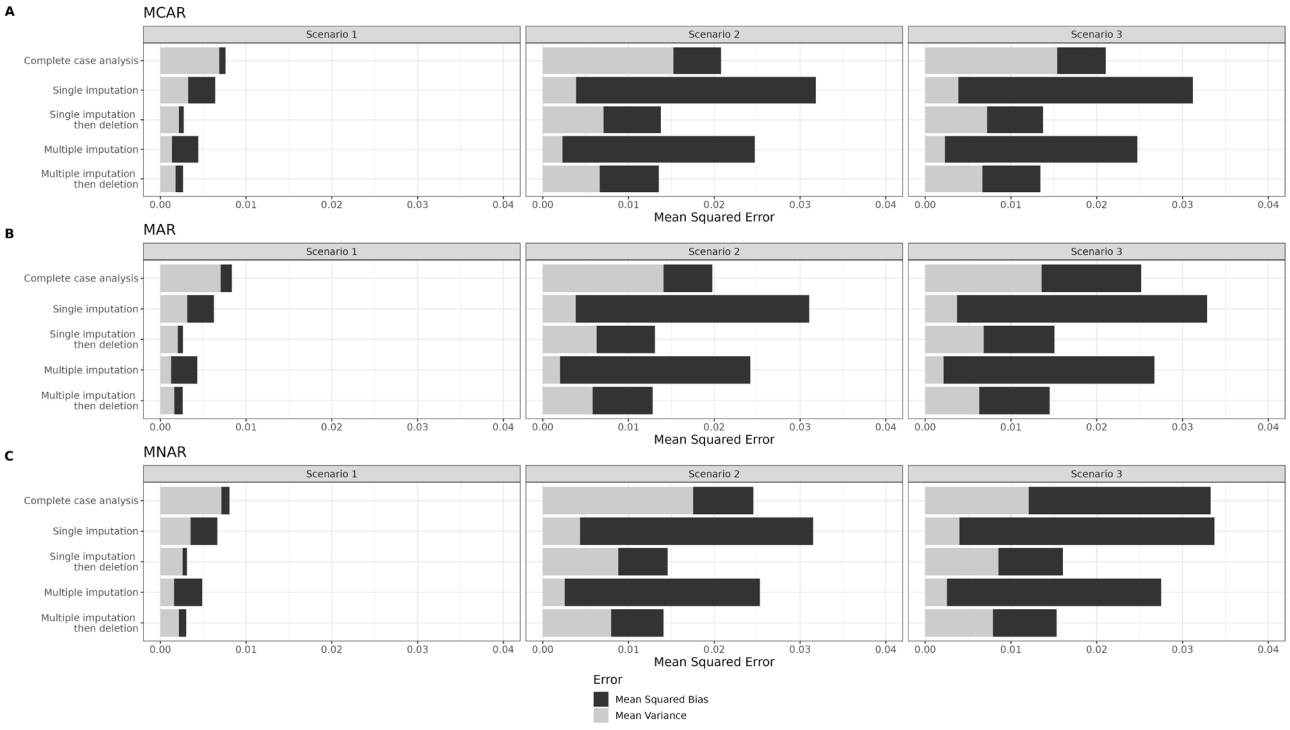

Figure 3 Mean squared error, mean squared bias and mean variance per method for each scenario under (A) MCAR, (B) MAR and (C) MNAR for a dichotomous outcome. Scenario 1: 10% of data missing in both case-mix (age, sex and baseline National Institutes of Health Stroke Scale (NIHSS)) and outcome (modified Rankin Scale, mRS) in all 17 hospitals. Scenario 2: 40% of data missing in the outcome and baseline NIHSS equally distributed over all hospitals. Scenario 3: 40% of data missing in the outcome and baseline NIHSS with the probability of missing data varying between hospitals. MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random.

For CCA, MID and SID, the MSE was mostly driven by variance under MCAR and MAR. By contrast, for SI and MI under MCAR and MAR, bias (ie, MSB) was the main driver of the MSE, more so for MI than for SI. Under MNAR, the MSE of all methods was mostly driven by bias. These results apply to both the ordinal and dichotomous outcomes. Hospitals with larger absolute reference effects had larger MSE for both outcomes, under all methods, mechanisms and scenarios (online supplemental figures 11 and 12).

Mean coverage of the 95% prediction intervals was highest for SID and MID, and was nearly 100% for SID and MID for nearly all mechanisms and scenarios for both outcomes (online supplemental tables 6 and 7; online supplemental figures 13 and 14). For the ordinal outcome, coverage dropped to around 90% for SID and MID under MNAR in scenarios 2 and 3. Coverage was near 100% for the other methods in scenario 1, but was considerably lower for the other methods for scenarios 2 and 3.

The Monte Carlo errors for the MSE and coverage were low. The maximum Monte Carlo errors for MSE and coverage for the ordinal outcome were 0.000436 and 0.0117, respectively, both for the CCA analysis in scenario 3 under MNAR. For the dichotomous outcome these were 0.000407 and 0.0115, respectively, also for the CCA analysis in scenario 3 under MNAR. See also online supplemental tables 8 and 9.

DISCUSSION

In this study, we used registry data to compare the bias and reliability of different methods of handling missing data when estimating hospital effects on outcome, assuming nine different patterns of missing data. We found that across these scenarios, the 'imputation, then deletion' method results in the least biased estimates of hospital effects for both ordinal and dichotomous outcomes.

Theoretically, under certain assumptions, both the CCA and imputation methods can result in valid estimates for coefficients in regression models. CCA is valid if maximum likelihood approaches are used (as in generalised linear mixed models), missing outcomes are MCAR or MAR and we condition our analysis on variables that govern missingness.[28 29] Imputation methods are valid if missing data are MCAR or MAR and if the imputation model is properly specified. This means that the imputation model should contain the same variables, interactions and non-linearities as the regression model in the main analysis (in our case the hospital effect estimation), as well as variables that govern missingness.[30] However, after SI, SEs (and consequently CIs and statistical inference) are invalid, as this method does not properly take into account uncertainty around the missing values. As a result, the SEs are underestimated. By contrast, MI does properly account for the uncertainty and leads to valid SEs.[29]

In our simulation study we found bias for each method to handle missing values, under every scenario and mechanism of missingness. These results can likely be explained by three reasons. First, while CCA is valid under an outcome that is MAR or MCAR and we properly specified the analysis model (the analysis was conditioned on the variables that governed missingness), mixed models apply shrinkage to the estimated random effects. Shrinkage improves estimation of the random effects by pulling them towards the mean, with the degree of shrinkage being influenced by the ratio of hospital-level variability versus residual variability, sample size per hospital and effect size.[31 32] Specifically, shrinkage is lower with higher variability on the hospital level, and hospitals with larger effect sizes and smaller sample sizes are shrunk more. As missing data reduce the sample size that can be used in the analysis, shrinkage will be larger and random effects biased. The effect of reduced sample size can also be seen with the increase in error of CCA from scenario 1 (34.39% of patients with missing data) to scenarios 2 and 3 (64% of patients with missing data).

Second, comparing the results from CCA with the standard imputation methods SI and MI, we would assume hospital effect estimates from imputation to be unbiased, as sample sizes are equal to the reference data set. However, error from the imputation methods is almost always larger than with CCA, and consists mostly of bias. This result could be explained by problematic imputations of the outcome.[16] In our imputation model, we used predictive mean matching to impute the ordinal outcome, because to our knowledge there is no imputation model that can handle clustered ordinal data. While predictive mean matching is robust against some model misspecification, using an imputation model that explicitly handles clustered ordinal data likely improves imputations of the outcome.[33] Predictive mean matching calculates a predicted value for the outcome from a linear mixed model, and then samples the outcome from a selected set of donors (patients without missing outcomes) with comparable predicted values. The imputation method used in this paper does not allow for drawing donors from only the same hospital. As the outcome can therefore be sampled from a different hospital, outcomes from hospitals will resemble each other more, leading to less hospital-level variability, increased shrinkage and thus more bias. However, the same results were found for the dichotomous outcome, for which the imputation model was correctly specified: a binomial generalised linear mixed model was used and all variables that govern missingness were included. A possible explanation is that shrinkage is also applied in the generalised linear mixed models used in the imputation procedure. This means that the imputed outcomes are shrunk towards the mean for each hospital, again leading to hospital effects that resemble each other

more with less hospital-level variability and thus more bias as a result.

Third, the bias under MNAR could be expected, as none of the methods for handling missing data are theoretically valid under MNAR. The possibility of data being MNAR should always be considered and, if suspected, sensitivity analyses should be performed.[34]

In this study, the least biased method for the estimated hospital effects on outcome is 'imputation, then deletion'. This method increases sample size by including patients with missing case-mix variables, while preventing the problematic imputations of the outcome. Concerning the precision of the hospital effects, we found that coverage of the 95% prediction intervals was highest for SID and MID for all mechanisms and patterns. Coverage was lower with increasing prevalence of missing data and MAR or MNAR mechanisms. Interestingly, coverage of the 95% prediction intervals was often higher than the nominal 95%. This might be due to the fact that the prediction intervals from the simulations were compared with the reference effect, which in turn was an estimation (ie, BLUP) that was subjected to shrinkage in the estimation procedure. We can infer that the prediction intervals for SID and MID have the highest probability of containing the hospital effect estimate from the complete data, but from our results we cannot make a statement about the coverage probability of the true and unknown underlying hospital effect. As we know from theory that the variance estimates arising from SI methods are invalid, for fixed effects at least, the MID approach would be preferred when aiming to estimate hospital effects on dichotomous and ordinal outcomes.[29]

Previous research has primarily focused on valid estimation of fixed effects in situations with missing data.[16 29] One study has looked at the bias of different MI methods when analysing incomplete data with a linear mixed effects model, and found that these validly estimated the variance of the random intercept distribution.[35] However, this study did not assess the bias and reliability of the random effects themselves, and only considered linear mixed models. We add to this literature by showing that random effects do not behave the same as fixed effects due to shrinkage in the estimation procedure for both ordinal and dichotomous outcomes.

However, some limitations of our research should be noted. First, this is a simulation study based on one data set. As such, results may not be directly generalisable to other settings, such as when outcomes or random effects have other distributions than those we have studied. While we aimed to provide a theoretical underpinning of our results, future studies should further investigate random effect estimation when data are missing. Second, we aimed to assess precision of the hospital effects, as measured with coverage of prediction intervals. We found that coverage of the prediction intervals was often higher than the nominal 95%, which might be due to the fact that these intervals were compared with an estimate from the reference data that was also subject to shrinkage. Future studies should evaluate precision using known reference effects to assess coverage of prediction intervals in the setting of missing data, for example, by first simulating data from a distribution with known parameters.

In conclusion, in case of missing data, estimates of hospital effects on ordinal and dichotomous outcomes are nearly always biased. The size of the bias is influenced by the proportion and mechanism of missing data. Reliability of the hospital effect estimates depends considerably on the method of imputation. The MID method seems the most promising method for handling missing data in terms of both bias and reliability.

Author affiliations
[1]Public Health, Erasmus MC, Rotterdam, Netherlands
[2]Gastroenterology and Hepatology, Franciscus Gasthuis en Vlietland, Rotterdam, Netherlands
[3]Department of Gastroenterology & Hepatology, Erasmus MC, Rotterdam, Netherlands
[4]Erasmus School of Health Policy and Management, Erasmus Universiteit Rotterdam, Rotterdam, Netherlands
[5]Neurology, Erasmus MC, Rotterdam, Netherlands
[6]Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, Netherlands
[7]Neurology, Amsterdam UMC, Amsterdam, Netherlands
[8]Neurology, Rijnstate Hospital, Arnhem, Netherlands
[9]Clinical Neurophysiology, University of Twente, Enschede, Netherlands
[10]Radiology and Nuclear Medicine, Amsterdam UMC, Amsterdam, Netherlands
[11]Radiology and Nuclear Medicine, MUMC+, Maastricht, Netherlands
[12]School for Mental Health and Sciences, Maastricht University, Maastricht, Netherlands

Collaborators MR CLEAN Registry Investigators Please see Supplementary File.

Contributors RCAvL, MA, NvL and FE designed the study. RCAvL and MA undertook analyses and interpretation of study findings and wrote the first draft of the manuscript. NvL, FE and HFL contributed to the interpretation of study findings, reviews, and revision of the manuscript. All the authors critically reviewed the various versions of the full paper and approved the final manuscript for submission. HFL is the guarantor of the study.

Funding This study was funded by Amsterdam University Medical Centers (no award/grant number), Erasmus Medisch Centrum (no award/grant number), TWIN Foundation (no award/grant number), Maastricht Universitair Medisch Centrum (no award/grant number).

Competing interests WvZ has received consulting fees from Codman and Stryker. DD has received research grants from Stryker and Bracco Imaging.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and the MR CLEAN Registry was approved by the Ethics Committee of the Erasmus University MC, Rotterdam, Netherlands (MEC-2014-235). With this approval, it was approved by the research board of each participating centre. At UMC Utrecht, approval to participate in the study has been obtained from their own research board and ethics committee. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. The data of the study cannot be made available to other researchers, as Dutch law prohibits data sharing when no patient approval was obtained for sharing coded data. However, syntax or output files of the statistical analyses may be made available for academic purposes upon reasonable request.

ORCID iDs
Reinier C A van Linschoten http://orcid.org/0000-0003-3052-596X
Marzyeh Amini http://orcid.org/0000-0002-0123-9510
Sanne J den Hartog http://orcid.org/0000-0002-1650-951X

REFERENCES

1 Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. *N Engl J Med* 2016;374:504–6.
2 Porter ME, Lee TH. The strategy that will fix health care. *Harv Bus Rev* 2013;91:1–19.
3 Kuhrij LS, Wouters MW, van den Berg-Vos RM, *et al*. The dutch acute stroke audit: benchmarking acute stroke care in the Netherlands. *Eur Stroke J* 2018;3:361–8.
4 Reeves MJ, Smith EE, Fonarow GC, *et al*. Variation and trends in the documentation of national institutes of health stroke scale in GWTG-stroke hospitals. *Circ Cardiovasc Qual Outcomes* 2015;8:S90–8.
5 Gale CP, Cattle BA, Moore J, *et al*. Impact of missing data on standardised mortality ratios for acute myocardial infarction: evidence from the myocardial ischaemia National audit project (MINAP) 2004-7. *Heart* 2011;97:1926–31.
6 Groene O, Kristensen S, Arah OA, *et al*. Feasibility of using administrative data to compare hospital performance in the EU. *Int J Qual Health Care* 2014;26 Suppl 1:108–15.
7 Kirkham JJ. A comparison of hospital performance with non-Ignorable missing covariates: an application to trauma care data. *Stat Med* 2008;27:5725–44.
8 Sterne JAC, White IR, Carlin JB, *et al*. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
9 Hutchings A, Neuburger J, van der Meulen J, *et al*. Estimating recruitment rates for routine use of patient reported outcome measures and the impact on provider comparisons. *BMC Health Serv Res* 2014;14:66.
10 Allison PD. *Missing data*. Thousand Oaks, CA: Sage Publications, n.d.: 1–12.
11 He Y. Missing data analysis using multiple imputation. *Circ Cardiovasc Qual Outcomes*2010;3:98–105.
12 Donders ART, van der Heijden GJMG, Stijnen T, *et al*. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59:1087–91.
13 Norris CM, Ghali WA, Knudtson ML, *et al*. Dealing with missing data in observational health care outcome analyses. *J Clin Epidemiol* 2000;53:377–83.
14 Raghunathan T. *Missing data analysis in practice*. Michigan: CRC Press, 2015.
15 Rubin DB. *Multiple imputation for nonresponse in surveys*. New York, US: Wiley, 1987.
16 von Hippel PT. Regression with missing ys: an improved strategy for analyzing multiply imputed data. *Sociol Methodol* 2007;37:83–117.
17 Jansen IGH, Mulder MJHL, Goldhoorn R-JB, *et al*. Endovascular treatment for acute ischaemic stroke in routine clinical practice: prospective, observational cohort study (MR CLEAN registry). *BMJ* 2018;360:k949.
18 van Swieten JC, Koudstaal PJ, Visser MC, *et al*. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* 1988;19:604–7.
19 Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials. *Stroke* 2007;38:1091–6.
20 Sullivan TR, Salter AB, Ryan P, *et al*. Bias and precision of the "multiple imputation, then deletion" method for dealing with missing outcome data. *Am J Epidemiol* 2015;182:528–34.
21 Amini M, van Leeuwen N, Eijkenaar F, *et al*. Improving quality of stroke care through benchmarking center performance: why focusing on outcomes is not enough. *BMC Health Serv Res* 2020;20:998.
22 van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011;45:1–67.
23 Christensen RHB. Ordinal---regression models for Ordinal data; 2022.
24 Bates D, Mächler M, Bolker BM, *et al*. Fitting linear mixed-effects models using Lme4. *J Stat Softw* 2015;67:1–48.
25 Holton GA. Value-at-risk: theory and practice. Available: https://www.value-at-risk.net/bias/ [Accessed 05 Jul 2023].
26 van Aert RCM, Schmid CH, Svensson D, *et al*. Study specific prediction intervals for Random-effects meta-analysis: a tutorial: prediction intervals in meta-analysis. *Res Synth Methods* 2021;12:429–47.
27 Koehler E, Brown E, Haneuse SJ-PA. On the assessment of Monte Carlo error in simulation-based statistical analyses. *Am Stat* 2009;63:155–62.
28 Beunckens C, Molenberghs G, Kenward MG. Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials. *Clin Trials* 2005;2:379–86.
29 Little RJA. Regression with missing X's: a review. *J Am Stat Assoc* 1992;87:1227.
30 van Buuren S. *Flexible imputation of missing data*. Boca Raton: CRC Press (Chapman & Hall), 2012.
31 Efron B, Morris C. Stein's paradox in statistics. *Sci Am* 1977;236:119–27.
32 Xu XS, Yuan M, Karlsson MO, *et al*. Shrinkage in nonlinear mixed-effects population models: quantification, influencing factors, and impact. *AAPS J* 2012;14:927–36.
33 Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol* 2014;14:75.
34 Resseguier N, Giorgi R, Paoletti X. Sensitivity analysis when data are missing not-at-random. *Epidemiology* 2011;22:282.
35 Huque MH, Moreno-Betancur M, Quartagno M, *et al*. Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model. *Biom J* 2020;62:444–66.