A comparison of incomplete-data methods for categorical data

van der Palm, D.W.; van der Ark, L.A.; Vermunt, J.K.

DOI
10.1177/0962280212465502
Publication date
2016
Document Version
Final published version
Published in
Statistical Methods in Medical Research

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2016, Vol. 25(2) 754–774

DOI: 10.1177/0962280212465502
smm.sagepub.com
SAGE

A comparison of incomplete-data methods for categorical data

Daniël W van der Palm, L Andries van der Ark and Jeroen K Vermunt

Abstract

We studied four methods for handling incomplete categorical data in statistical modeling: (1) maximum likelihood estimation of the statistical model with incomplete data, (2) multiple imputation using a loglinear model, (3) multiple imputation using a latent class model, (4) and multivariate imputation by chained equations. Each method has advantages and disadvantages, and it is unknown which method should be recommended to practitioners. We reviewed the merits of each method and investigated their effect on the bias and stability of parameter estimates and bias of the standard errors. We found that multiple imputation using a latent class model with many latent classes was the most promising method for handling incomplete categorical data, especially when the number of variables used in the imputation model is large.

Keywords

Missing data, categorical data, multiple imputation, latent class analysis, MICE, maximum likelihood, medical research

1 Introduction

This paper discusses methods to handle incomplete categorical data. Many medical studies deal solely with analyzing categorical data and, consequently, the statistical model that is used to analyze the data (from here on referred to as the *substantive model*) is also tailored to categorical data. For example, predictors of reduced length of hospital stay were studied using logistic regression,[1] determinants of caregivers' health were studied using loglinear modeling,[2] and the effectiveness of the World Health Organization Disability Assessment Schedule II was investigated using a nonparametric item response analysis.[3] A frequently encountered problem is that the data are incomplete, which prevents a straightforward statistical analysis; a researcher should handle this problem appropriately. Klebanoff and Cole[4] found that the majority of applied researchers resort to ad-hoc methods such as complete-case analysis or pair-wise deletion, which may lead to biased statistical results[5] and reduced power.[5,6]

Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands

Corresponding author:
Daniël W van der Palm, Tilburg School of Social and Behavioral Sciences, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands.
Email: d.w.vdrpalm@uvt.nl

For handling incomplete continuous data, adequate alternatives have been proposed, extensively researched,[7] and implemented in major software packages such as SPSS[8] and SAS.[9] Hence, there is no need for applied researchers to resort to ad-hoc methods in case of continuous data.

Incomplete data methods for categorical data have not yet been crystallized out, and it is unknown which method should be recommended to practitioners. Ideally, an incomplete-data method should meet three criteria. For the substantive model, it should produce parameter estimates (i) that are unbiased, (ii) that are stable in order to avoid unnecessary loss of power in the statistical analysis, and (iii) that have standard errors correctly reflecting the uncertainty due to missing data. Ideally, these criteria should be met for data sets with both small and large numbers of variables, sample sizes, and percentages of incomplete data, and for both simple and complex associations in the data.

With respect to these criteria, two incomplete-data methods for categorical data are especially promising: *Multiple imputation using latent class analysis* (MILC[10,11]) and *multivariate imputation using chained equations* (MICE[12–14]). Both methods have the practical advantage that they can easily handle data sets containing a large number of variables and respondents. However, researchers having incomplete categorical data cannot yet readily apply MILC and MICE because there are various unresolved issues (explained hereunder). The impact of these issues on the three criteria for substantive models is unknown. In this study, we discuss two reasonable options for the unresolved issues for both MILC and MICE, and investigate to which degree they meet the three criteria, so as to decide which incomplete-data method should be selected for categorical data. *Multiple imputation using a loglinear model* (MILL[6]) and *maximum likelihood for incomplete data* (MLID,[5,15–17] also known as full information maximum likelihood) are used as benchmarks. MILL is known to produce unbiased parameter estimates[6,18,19] but can only handle a small number of variables; MLID is known to be asymptotically unbiased but may run into difficulties as the number of variables becomes very large.[10]

The remainder of this paper is organized as follows. First, we briefly discuss the four incomplete-data methods. For both MILC and MICE, we discuss two variants, resulting in six incomplete-data methods in total. Second, we compare the advantages and disadvantages of the methods in a theoretical discussion. Third, we present the results of two simulation studies. In Study 1, for dichotomous data, we compared MILC, MICE, MILL, and MLID with respect to the three criteria. In Study 2, for trichotomous data, we compared MILC, MICE, and complete-case analysis with respect to the three criteria. Fourth, we applied MLID, MILC, MICE, and complete-case analysis to a medical data set. Finally, we give recommendations based on the theoretical discussion and the two simulation studies.

2 Incomplete-data methods

2.1 Incomplete data

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_J)$ denote the scores on the J variables, and let $\boldsymbol{\theta}$ be the generic notation for the vector of unknown parameters of the joint distribution of $\mathbf{Y}$, denoted $P(\mathbf{Y}; \boldsymbol{\theta})$. To distinguish specific models Greek letters other than $\boldsymbol{\theta}$ may also be used to denote paramete r vectors. Note that Y_j may be either a predictor variable or an outcome variable depending on the substantive model. If confusion arises, we add the superscripts p and o to indicate that a variable serves as a predictor variable or outcome variable, respectively. $\mathbf{Y}$ may contain missing values, and the objective is to deal with them appropriately.

Most incomplete-data methods, including the ones considered in this paper, assume that the mechanism that caused the missing values is *ignorable*,[16] which means that two conditions should

hold. First, the parameters that govern the missing data process must be unrelated to the parameters to be estimated, which is a rather unrestrictive assumption.[5] Second, the data must be *missing at random* (MAR), which means that whether or not a score is missing only depends on scores observed in the study. If, after conditioning on all observed data, the missingness depends on missing values of variables included in the study or on variables not included in the study, MAR is violated and, as a result, the missingness mechanism is non-ignorable. Non-ignorable missingness may cause biased parameters in the substantive model (first criterion). Apart from special studies with planned missingness,[6] MAR is unlikely to hold in practice, and it is impossible to test whether the MAR assumption holds for a particular data set.[7] Therefore, the degree to which MAR is violated (i.e. the degree to which the observed scores cannot explain the missingness mechanism) becomes important: If the violation of MAR becomes more severe, the parameter bias in the substantive model is likely to increase. If the number of variables in a data set increases, the degree to which the variables can explain the missingness mechanism is also likely to increase. Hence, if an incomplete-data method can handle a large number of variables, and if a large number of variables is available, the violation of MAR will most likely be less severe and the missingness mechanism is more likely to be ignorable. This notion[6] plays an important role in our evaluation of incomplete-data methods and will be referred to as *Schafer's notion on the number of variables*.

2.2 Description of incomplete-data methods

2.2.1 Maximum likelihood for incomplete data

MLID is a well-known and documented method to obtain parameter estimates and standard errors in the presence of missing data.[5] MLID constitutes estimating the parameters of the substantive model and their standard errors, using all observed data. For example, when studying predictors of reduced length of hospital stay using logistic regression,[1] MLID can be used to estimate the logistic regression model using all observed data. No further action is required; the obtained parameter estimates and standard errors can be directly interpreted. The substantive model can be an asymmetric model such as a logistic regression model or an item response theory model, which describe the conditional distribution of the outcome variables given the predictor variables $P(\mathbf{Y}^o|\mathbf{Y}^p;\boldsymbol{\theta})$, or a symmetric model, such as a loglinear model, latent class model, or canonical correlation model, which describe the joint distribution of all variables $P(\mathbf{Y};\boldsymbol{\theta})$. MLID assumes that the missingness mechanism is ignorable. For categorical data, specialized software is usually required to conduct MLID, such as LEM[20] or Mplus.[21]

2.2.2 Multiple imputation

Multiple imputation consists of creating m completed data sets by replacing the missing values in the data with plausible values m times. These plausible values replacing the missing values are called the imputed values. The statistical model that generates imputed values is referred to as the *imputation model*. After the multiple imputation, on each of the m completed data sets a substantive model is estimated, and the m sets of parameter estimates and standard errors are combined into a single set. Most researchers use $m = 5$, but this value is currently debated.[22] Using multiple imputation allows for separating the missing data handling and the substantive analysis; a researcher can estimate substantive models as if there had been no missing data, or distribute the completed data to other researchers for further analysis.

Multiple imputation starts in the same way as MLID for symmetric models: A statistical model is estimated describing the joint distribution $P(\mathbf{Y};\boldsymbol{\theta})$. Rather than a substantive model, this model is an imputation model for obtaining imputed values from $P(\mathbf{Y};\boldsymbol{\theta})$. For example, when studying

predictors of reduced length of hospital stay using logistic regression,[1] a loglinear model describing the joint distribution of both predictor variables and reduced length of hospital stay may be used as an imputation model to generate imputed values replacing the missing data m times. After the multiple imputation, logistic regression analysis can be conducted on the completed data sets.

One must account for the fact that the imputed values are not observed and, therefore, uncertain. There are two sources of uncertainty.[23] Firstly, the estimated parameters of the imputation model are uncertain; this uncertainty is expressed by their standard errors. Secondly, there is uncertainty due to sampling variability when drawing imputed values from $P(\mathbf{Y}; \boldsymbol{\theta})$. To account for parameter uncertainty, for each of the m data sets, a different set of parameters of the imputation model is used. In a Bayesian framework, the m sets of parameters of the imputation models are random draws from $P(\boldsymbol{\theta}|\mathbf{Y})$, the distribution of the parameters given the data.[5] In a frequentist framework, the m sets of parameters are estimated using m nonparametric bootstrap samples of the data.[10] A nonparametric bootstrap sample consists of randomly drawing a new sample of N observations with replacement.[24] To reflect uncertainty due to sampling variability, the replacement of missing values is done m times, yielding m completed data sets. The three multiple imputation methods for categorical data discussed in this paper differ in the way that they describe the joint distribution, $P(\mathbf{Y}; \boldsymbol{\theta})$, and how they account for parameter uncertainty. MILL is discussed briefly because this method is ready for use; MILC and MICE are described in more detail so as to allow the discussion of the specific options these methods offer.

Multiple imputation using a loglinear model. MILL uses a loglinear model as the imputation model. Let the parameters of the loglinear model be denoted $\boldsymbol{\lambda}$; the saturated loglinear model for dichotomous responses can be written as

$$\log P(\mathbf{Y}; \boldsymbol{\lambda}) = \lambda + \sum_{i=1}^{J} \lambda_i Y_i + \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \lambda_{ij} Y_i Y_j + \cdots + \lambda_{1,2,\ldots,J} Y_1 Y_2 \cdots Y_J. \quad (1)$$

The joint distribution is obtained by taking the exponential of the right-hand side of equation (1). Typically, a saturated loglinear model is used to obtain imputation values because it captures all possible associations in the data; therefore, it is the gold standard for multiple imputation of categorical data.[10] If higher-order interaction terms are omitted, the approximation of the joint distribution by the loglinear model may deteriorate. MILL can, for example, be conducted using software packages CAT[6] or Latent GOLD 4.5,[25] which utilize a Bayesian and a nonparametric bootstrap approach, respectively, to account for parameter uncertainty.

Multiple imputation using a latent class model. MILC uses a latent class model to estimate the joint distribution of the variables in the data. Let X denote a discrete latent variable with K latent classes, indexed by k ($k = 1, \ldots, K$). Let $\boldsymbol{\pi}$ denote the vector of parameters of the latent class model; $\boldsymbol{\pi}$ can be divided into $\boldsymbol{\pi}_x$, the latent class proportions, and $\boldsymbol{\pi}_y$ the conditional response probabilities. Under a latent class model, joint distribution $P(\mathbf{Y}; \boldsymbol{\pi})$ has the following form[26–29]:

$$\begin{aligned} P(\mathbf{Y}; \boldsymbol{\pi}) &= \sum_{k=1}^{K} P(X = k; \boldsymbol{\pi}_x)\, P\big(\mathbf{Y}|X = k; \boldsymbol{\pi}_y\big) \\ &= \sum_{k=1}^{K} P(X = k; \boldsymbol{\pi}_x) \prod_{j=1}^{J} P\big(Y_j|X = k; \boldsymbol{\pi}_{y_j}\big). \end{aligned}$$

If the number of latent classes is sufficiently large, a latent class model correctly picks up the first, second, and higher order moments of the response variables, as is the case with all forms of mixture models.[30] It is unknown how many latent classes are sufficient for a good approximation of the joint distribution. Vermunt et al.[10] argued that it is better to have too many than too few latent classes. Therefore, out of three selection criteria, Akaike's information criterion (AIC[31]), Bayesian information criterion (BIC[32]), and AIC3,[33] they suggested using AIC to select the number of latent classes because it yields the largest number of latent classes. Hence, letting AIC determine the number of latent classes, abbreviated *MILC (AIC)*, is the first option for MILC. However, it is expected that an even larger number of latent classes can further improve the approximation of the joint distribution. Having a relatively large number of latent classes, abbreviated *MILC (Large)*, is the second option for MILC. MILC can be applied using Latent GOLD,[25] which uses the nonparametric bootstrap to account for parameter uncertainty.

Multivariate imputation using chained equations. MICE[12] is a fully conditional specification method, which specifies the imputation model on a variable-by-variable basis using a separate conditional distribution for each incomplete variable. Let $\mathbf{Y}_{-j}$ denote the scores on all variables except Y_j. MICE reduces the problem of finding one J-dimensional joint distribution $P(\mathbf{Y};\boldsymbol{\theta})$ to finding J univariate conditional distributions $P(Y_1|\mathbf{Y}_{-1};\boldsymbol{\theta}),\ldots,P(Y_J|\mathbf{Y}_{-J};\boldsymbol{\theta})$.[12-14] Conditional distribution $P(Y_j|\mathbf{Y}_{-j};\boldsymbol{\theta})$ is used for imputation of Y_j $(j=1,\ldots,J)$. Under certain conditions, a draw from each of the J conditional distributions is equivalent to a single draw from the joint distribution,[14] but it is not guaranteed. Results from simulation studies[13,34] suggest that the problem is unlikely to be serious in practice.

MICE starts with replacing missing values of the variables by draws from their respective marginal distributions. Next, in an iterative process, the imputed values are updated variable by variable using the univariate conditional distributions. When Y_j is imputed, the other variables act as predictor. If the joint distribution that is defined by the J conditional distributions exists then this iterative process is a Gibbs sampler[14] and converges to the joint distribution of the J variables. Often, as little as 10 to 20 iterations are required.

The imputation model describing the conditional probabilities $P(Y_1|\mathbf{Y}_{-1};\boldsymbol{\theta}),\ldots,P(Y_J|\mathbf{Y}_{-J};\boldsymbol{\theta})$ can be any appropriate regression model depending on the nature of the outcome variable[35]: linear regression in combination with predictive mean matching, logistic regression, polytomous regression, and nonlinear regression. We focused on two imputation models; the first one being logistic regression (abbreviated MICE (LOG)) which is the default method in the R-package MICE[12] for dichotomous outcome variables (for Study 2, polytomous regression is used, which is the extension of MICE (LOG) to variables with more than 2 categories; for details see, e.g. Van Buuren et al.[12]). Let $\boldsymbol{\beta}$ denote the vector of parameters for the logistic regression model. MICE (LOG) models conditional distribution $P(Y_j|\mathbf{Y}_{-j};\boldsymbol{\beta})$ as

$$\text{logit}\left[P\left(Y_j|\mathbf{Y}_{-j};\boldsymbol{\beta}\right)\right]=\beta_0+\beta_1Y_1+\cdots+\beta_{j-1}Y_{j-1}+\beta_{j+1}Y_{j+1}+\cdots+\beta_JY_J$$

We also considered linear regression in combination with predictive mean matching (abbreviated MICE (PMM)). The first step of MICE (PMM) is to obtain a predicted value by means of linear regression in which all other variables serve as predictors. In the second step, the respondent that has the most similar predicted value as well as an observed value on the variable that is being imputed is selected as the nearest neighbor. Subsequently, the observed value of this nearest neighbor is used as the imputation value for the respondent with a missing value.

Parameter uncertainty is accounted for in a Bayesian framework; a new set of parameters is drawn from its posterior distribution for the construction of each imputed data set. More specifically, the MICE algorithm involves iteratively sampling parameter values $\boldsymbol{\beta}$ from their posterior distribution and imputing the missing values Y_j by drawing from the conditional distribution $P(Y_j|\mathbf{Y}_{-j};\boldsymbol{\beta})$. This corresponds with a Gibbs sampling scheme if the joint distribution of the variables can be constructed from their univariate conditional distributions and if the distribution from which parameters are drawn can be constructed from the joint distribution of the variables and an appropriate prior distribution.[12] These two conditions are not fulfilled when using MICE with categorical data, which means that the algorithm is not an exact Gibbs sampler. MICE can be conducted using the R package MICE[12] or the STATA[36] package ICE.[22,37]

2.2.3 Other incomplete-data methods

We have three remarks on other incomplete-data methods. First, besides MLID and multiple imputation, there are two other categories of incomplete-data methods: the *fully Bayesian* method[38] and *weighted estimating equations*.[16] We did not consider these two approaches to limit the scope of this paper. A full Bayesian analysis with for example WinBugs is in fact similar to both MLID and multiple imputation; that is, the parameters of the substantive model are estimated using the incomplete data using an algorithm containing a step in which the missing values are imputed.[38] Results can be expected to be similar to MLID. Weighting is typically used to deal with completely missing data and has limited practical use with partially missing data.[39] It may moreover yield instable estimates in the presence of influential weights.[40]

Second, a popular imputation model for multiple imputation is the multivariate normal distribution.[7] The method is robust against deviations from normality[41] and may even perform well for categorical data,[42] although some studies reported serious bias.[43,44] We did not consider incomplete data-methods that were not designed for categorical data as these methods are not suitable for nominal variables (e.g., blood type, eye color, surgical outcome).

Third, the best known ad-hoc method is probably complete-case analysis, in which only the observations without any missing values are used to estimate the substantive model. In other words, subjects who have at least one missing value are discarded from the analysis. Hence, in contrast to MLID, complete-case analysis does not incorporate all available information. Complete-case analysis reduces power and may yield biased parameter estimates for the substantive model if the data are not missing completely at random (MCAR)[5]; this MCAR assumption is considered to be unrealistic in most situations.[7] Complete-case analysis is included in Study 2 and the real-data example. For Study 2, the number of variables was too large for more preferable benchmarks such as MILL and MLID.

2.3 Advantages, disadvantages, and unresolved issues of the incomplete-data methods

2.3.1 Practical issues

For application of the incomplete-data methods, sample size, complexity of the association structure in the data, and percentage of missingness are not restrictive for any of the methods. A limitation of MILL is that it cannot handle large numbers of variables because the number of cells in the contingency table that has to be evaluated in the loglinear model becomes too large. For example, the number of cells that need be evaluated exceeds one million for 20 dichotomous variables and one billion for 30 dichotomous variables, 19 trichotomous variables, or 13 variables with five categories. In cases where the substantive model contains fewer variables than available

in the data set, a possible solution is to consider only those variables that are used in the substantive model. However, following Schafer's notion on the number of variables, using only a small number of variables for the imputation model may result in biased parameter estimates. For MILC and MICE, large numbers of variables do not pose a problem. A potential problem for MLID is that it usually requires specialized software, depending on the substantive model that one wants to estimate, whereas standard data-analysis techniques can be applied after the imputation phase of MILL, MILC and MICE. Moreover, MLID can only be used if the number of variables in substantive model is not too large.

2.3.2 Bias

We consider three possible causes of bias in the parameter estimates: First, non-ignorable missingness in the data. Following Schafer's notion on the number of variables, it is suggested that the inclusion of many variables in the imputation model makes it more likely that violations of ignorability are minor. The second possible cause of bias is misspecification of the imputation model so that it is too parsimonious. The imputation model should be as general as possible; this ensures that the imputed values behave as neutral as possible in subsequent analyses.[6] Hence, the main criterion of an adequate imputation model is whether it captures all the associations between categorical variables that exist on the population level.[6] The third possible cause of bias is misspecification of the substantive model. However, this is unrelated to the incomplete-data method being used and is not pursued further.

For MLID, no imputation model needs to be specified but a violation of the ignorability assumption may result in biased parameter estimates. Statistical analyses that are based on MLID only include those variables in the data that are substantively relevant, possibly excluding many variables. When the number of variables in the substantive model is small, then, following Schafer's notion on the number of variables, the missingness mechanism in the reduced data is less likely to be ignorable. Simulation studies showed that under ignorable missingness, MLID yields unbiased parameter estimates.[6]

For MILL, the imputation model being too parsimonious is not an issue because the imputation model is typically the saturated model. However, MILL can handle only a limited number of categorical variables. As a result, following Schafer's notation on the number of variables, the missingness mechanism in the reduced data may not be treated as ignorable possibly resulting in biased parameter estimates. Simulation studies showed that under ignorable missingness, MILL yields unbiased parameter estimates.[6]

For MILC and MICE, the amount of non-ignorable missingness may be reduced if the data contain many variables relevant for predicting the missing values (Schafer's notion on the number of variables) because both methods can handle a very large number of (auxiliary) variables. For MICE, it is unknown which of the two variants yields the least bias, for MILC, it is expected that a large number of latent classes, MILC (Large), produces less bias than a smaller number of latent classes, MILC (AIC).

2.3.3 Stability

We consider three possible causes that influence the stability of parameter estimates in the presence of incomplete data. A first possible cause is a too small effective sample. It is well known sample size has a positive effect on stability.[45] None of the incomplete-data methods under investigation unduly reduce the effective sample size, in the way some ad hoc methods do (e.g., complete-case analysis, pair-wise deletion). However, it is unknown whether the incomplete-data methods under investigation yield the same stability of parameter estimates given a fixed sample size. A second

possible cause is misspecification of the imputation model so that it is too complex. This is the well-known tradeoff between bias and stability: If the imputation model is too parsimonious it may result in biased outcomes, if it is too complex, it may result in less stable outcomes. For most researchers, unbiased parameter estimates are more important than stable parameter estimates. The third possible cause of instability is misspecification of the substantive model so that it is too complex. However, this is unrelated to the incomplete-data method being used and is not pursued further.

Only for the second possible cause of instability, an overly complex imputation model, we have expectations for the incomplete-data methods under investigation. MLID does not require an imputation model, so no loss of stability can ensue from an overly complex imputation model. For MILL, the imputation model is saturated meaning that it is expected to be overly complex in most cases. Therefore, a certain loss of stability is expected for MILL in comparison to MLID.

For MILC, the two variants are expected to differ in stability because their respective imputation models differ in complexity. MILC (Large) uses a relatively large number of latent classes which means that its imputation model is expected to be able to capture every possible association. As is the case with MILL, results produced by MILC (Large) are expected to lose a certain degree of stability because its imputation model is expected to be overly complex. MILC (AIC) estimates the required number of latent classes using AIC, which results in a relatively small number of latent classes. Therefore, its imputation model is more parsimonious and its results are expected to be more stable than MILC (Large).

For MICE, the two variants differ in the conditional imputation model that is used. The stability of MICE depends on the degree to which higher order associations are included in the conditional imputation model. The default setting of MICE (PMM) only includes main effects. However, because predictive mean matching is used, all associations can be picked up for data sets with a small number of variables. Therefore, we expect that the stability of the parameter estimates produced by MICE (PMM) is similar to MILL and MILC (Large). The default setting of MICE (LOG) also only includes main effects. Therefore, MICE (LOG) is expected to have relatively stable results.

2.3.4 Bias of the standard errors

It is unknown whether the six incomplete-data methods overestimate or underestimate the standard errors of parameter estimates. Hence, we have no specific expectations with regard to the bias of the standard errors.

3 Study 1: Bias, stability, and bias in standard errors produced by MILC, MICE, MILL, and MLID for a small number of dichotomous variables

In Study 1, we compared incomplete-data methods MILC (AIC), MILC (Large), MICE (PMM), and MICE (LOG) to MLID and MILL, on bias of the parameter estimates, stability of the parameter estimates, and bias of the standard errors. Because MLID and MILL can handle only a limited number of variables, the number of variables was kept small. The design of Study 1 was motivated by the study of Kurian et al.,[1] who studied several predictors of a single outcome variable reduced ''length of hospital stay'' using logistic regression.

3.1 Method

3.1.1 General set up

The set up of the simulation study was as follows. First, we sampled complete data sets from a population model. Second, we created incomplete data sets by deleting variable scores according to

an MAR missingness mechanism. Third, for each incomplete data set we constructed five completed data sets using a missing-data method. Fourth, we used the completed data sets to estimate the parameters of the regression part of the population model and we reported the bias and stability of the parameter estimates.

The population model was defined for five dichotomous predictor variables $Y_1, \ldots, Y_5$, and one dichotomous outcome variable Y_6. The categories were coded 0 and 1 (dummy coding). Dummy coding was used because it is the most commonly used coding scheme for logistic regression models. The associations among the predictor variables $Y_1, \ldots, Y_5$ were described by loglinear model

$$\log P(Y_1, Y_2, Y_3, Y_4, Y_5) = -1.47 + \sum_{j=1}^{5} -2 \cdot Y_j + \sum_{j=1}^{4} \sum_{k=j+1}^{5} 1 \cdot Y_j Y_k \tag{2}$$

Outcome variable Y_6 was related to the predictor variables by logit model

$$\text{logit}(Y_6) = \beta_0 + Y_1 + \beta_2 \cdot Y_2 + \beta_3 \cdot Y_3 + Y_4 + Y_5 - \beta_{23} \cdot Y_2 \cdot Y_3, \tag{3}$$

which contains main effects of the predictor variables as well as the interaction effect of Y_2 and Y_3. The strength of the interaction term, β_{23}, was manipulated in the study. The coefficients β_0,β_2 and β_3 are changed together with β_{23} so that the average logit and the average effects of Y_2 and Y_3 remain constant across conditions. Complete data sets were created by sampling from $P(Y_1, Y_2, Y_3, Y_4, Y_5)$ (equation (2)) and $P(Y_6|Y_1, Y_2, Y_3, Y_4, Y_5)$ (equation (3)).

Variables Y_1 and Y_2 had missing values that were MAR. Variables R_1 and R_2 indicated whether a score was missing, $R_i = 0$, or observed, $R_i = 1$, for Y_1 and Y_2, respectively. Missing values in Y_1 were created using logistic regression model

$$\text{logit}(R_1) = \gamma_1 + 1.09 \cdot Y_3 + 2.01 \cdot Y_4 - .79 \cdot Y_3 Y_4, \tag{4}$$

and missing values in Y_2 were created using logistic regression model

$$\text{logit}(R_2) = \gamma_2 + 1.04 \cdot Y_5 + 1.94 \cdot Y_6 - .74 \cdot Y_5 Y_6 \tag{5}$$

The total percentage of missingness (one of the predictor variables in Study 1, to be discussed later) was manipulated by changing the intercepts (γ_1 and γ_2) in equations (4) and (5). This approach allows for varying the total percentage of missingness without altering the strength of associations between the predictor variables and the missingness indicator variable in equations (4) and (5).

For each incomplete data set, five completed data sets were created using a multiple imputation method and for each completed data set logistic regression model

$$\text{logit}(Y_6) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \beta_5 Y_5 + \beta_{23} Y_2 Y_3 \tag{6}$$

was estimated. Rubin's[23] rules were used to combine the five sets of regression parameter estimates. It should be noted that $m = 5$ completed data sets is usually considered to be sufficient to obtain stable results.[6] However, other researchers have argued that m should be based on the total percentage of missingness; for example, m should equal the total percentage of missing data to obtain a sufficient degree of stability in the results.[22] In many cases, this would render m larger than five. We note that this is especially important for a single analysis; in a simulation study the size of m has much less influence because of the large number of replications.

Three software packages were used for multiple imputation and parameter estimation. Data were generated using software package LEM,[20] methods MILC and MILL were conducted using the software program Latent GOLD 4.5,[29] and for MICE we used the R package MICE.[12] After multiple imputation, the substantive model, defined in equation (6), was estimated using Latent GOLD 4.5. For MLID, the substantive model was estimated using LEM.

3.1.2 Predictor variables

Incomplete-data method was a within factor with six levels: MILC (AIC), MILC (Large), MICE (PMM), MICE (LOG), MLID, and MILL. The incomplete-data methods determine the imputation model and may thus affect both bias and stability.

Strength of the interaction term was a between factor with three levels that was manipulated by varying parameter β_{23} in equation (3). The levels were no three-way association ($\beta_{23} = .00$), medium ($\beta_{23} = -.80$), and strong ($\beta_{23} = -2.00$). Strength of the three-way association sets requirements for the complexity of the imputation model. If this effect increases, a more complex imputation model is required to pick up the associations in the data. It is expected that strength of the three-variable association affects both bias and stability.

Percentage of missingness was a between factor with three levels: moderate (10% missingness), high (20% missingness), and extreme (40% missingness). The percentage of missingness was manipulated by varying parameters γ_1 and γ_2 in equations (4) and (5), respectively. For 10% missingness, $\gamma_1 = -2.46$ and $\gamma_2 = -2.53$, for 20% missingness, $\gamma_1 = -1.41$ and $\gamma_2 = -1.44$, and for 40% missingness, $\gamma_1 = -.39$ and $\gamma_2 = -.41$. As the percentage of missingness increases, the imputation model becomes more important. The condition with 40% of missingness is included because the consequences of an inadequate imputation model are magnified by an increase in the percentage of missingness.

Sample size was a between factor with two levels: Small ($N = 200$) and large ($N = 1000$). Sample size is expected to predominantly affect stability. In particular, the aim was to examine how sample size is related to the stability of the statistical results in the analysis of interest for each missing-data method.

The four predictor variables were fully crossed producing a $6 \times 3 \times 3 \times 2$ design, with 1000 replications for each of the 18 combinations of the between-subjects variables.

3.1.3 Outcome variables

The outcome variables were bias of parameter estimates, standard deviation of parameter estimates across replications, and bias of the reported standard errors.[7,45] Let $\hat{\beta}_{bj}$ denote a parameter estimate of the jth variable (equation (6)) in replication b ($b = 1, \ldots, 1000$), then the bias over 1000 replications was computed as

$$\text{bias} = \frac{1}{1000} \sum_{b=1}^{1000} \left(\hat{\beta}_{bj} - \beta_j \right).$$

Stability, denoted by $sd(\hat{\beta}_j)$, was measured by the standard deviation of parameter estimates across replications and was computed as

$$\text{sd}(\hat{\beta}_j) = \sqrt{\frac{1}{999} \sum_{b=1}^{1000} \hat{\beta}_{bj}^2}\ .$$

Let $\text{se}(\hat{\beta}_{bj})$ denote the estimated standard error of parameter estimate $\hat{\beta}_{bj}$. Bias of the reported standard errors (BSE) was computed as

$$\text{BSE} = \frac{1}{1000}\sum_{b=1}^{1000}\left[\text{se}\left(\hat{\beta}_{bj}\right) - \text{sd}\left(\hat{\beta}_{j}\right)\right],$$

The bias and stability of parameter estimates and bias of the standard errors were only considered for parameters β_2 (a main effect that is influenced by the interaction effect β_{23}), β_4 (a main effect that is not influenced by the interaction effect β_{23}), and the interaction effect β_{23}.

3.2 Results

3.2.1 Bias

Table 1 shows the most important results for bias for 40% missingness. For lower percentages of missingness, the pattern of the bias was similar but the absolute values were smaller. This confirms that for larger percentages of missingness, the imputation model becomes more important. The most important result is that incomplete-data methods MLID, MILC (AIC), and MICE (LOG) produced

Table 1. Bias in the estimates of three logistic regression coefficients for six incomplete-data methods, three different levels of strength of three-variable associations ($\beta_{23} = 0$, $\beta_{23} = -.8$, $\beta_{23} = -2$), two sample sizes (200, 1000), and 40% missingness.[a]

			Incomplete-data method					
N	β_{23}	RC	MLID	MILL	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
200	0	$\beta_2 = 1$	.040	.076	.082	.025	.051	.073
		$\beta_4 = 1$	.065	.067	.061	.041	.068	.061
		$\beta_{23} = 0$	.046	−.014	−.008	.031	.014	.009
	−.8	$\beta_2 = 1.4$	.050	.094	.086	.040	.058	.058
		$\beta_4 = 1$	.071	.057	.058	.042	.062	.057
		$\beta_{23} = -.8$	−.036	.010	.022	**.258**	.036	**.310**
	−2	$\beta_2 = 2$	**.091**	**.103**	.057	**−.241**	.057	**−.357**
		$\beta_4 = 1$	.058	.040	.036	.002	.055	.021
		$\beta_{23} = -2$	**−.144**	−.074	−.006	**.541**	−.025	**.733**
1000	0	$\beta_2 = 1$	−.005	.010	.009	−.003	.011	.008
		$\beta_4 = 1$	.015	.013	.015	.012	.016	.014
		$\beta_{23} = 0$	−.019	−.004	−.006	−.001	−.006	−.004
	−.8	$\beta_2 = 1.4$	.026	.036	.036	**−.088**	.036	**−.136**
		$\beta_4 = 1$	.005	.003	.004	−.004	.003	−.002
		$\beta_{23} = -.8$	−.010	−.016	−.015	**.205**	−.014	**.302**
	−2	$\beta_2 = 2$	.006	.033	.028	**−.216**	.031	**−.410**
		$\beta_4 = 1$	.014	.011	.011	−.015	.012	−.013
		$\beta_{23} = -2$	−.036	−.046	−.041	**.457**	−.056	**.765**

Note: N: sample size; β_{23}: strength of three-variable association; RC: regression coefficient. For MILC (AIC) the average number of classes indicated by AIC ranged from 2.8 ($N = 200$, $\beta_{23} = -2$) to 3.8 ($N = 1000$, $\beta_{23} = -2$), for MILC (Large) a constant number of 12 classes was used.

[a]Remarkable bias is printed in boldface.

large bias in the estimates of β_2 and β_{23} when there was an interaction effect in the data ($\beta_{23} \neq 0$), whereas estimates of β_4, a parameter not influenced by an interaction effect, showed much less bias. These results suggest that MILC (AIC) and MICE (LOG) have imputation models that are too parsimonious to pick up the three-way association in the data. Furthermore, the results suggest that MLID cannot handle very well the combination of a small sample size and a complex association. Seemingly, the asymptotic property of unbiased parameter estimates is not fully established under these circumstances. A second result is that MILL, which we used as a gold standard, produced similar bias or sometimes more bias (e.g., for β_2 in condition $\beta_{23} = -2$, $N = 200$) than MILC (Large) and MICE (PMM). A third result is that the bias was slightly larger for $N = 200$ than for $N = 1000$, which indicates that increased sampling variability may somewhat increase bias.

3.2.2 Stability

Table 2 shows the most important results for stability for 40% missingness. The most important result is that stability does not change dramatically across incomplete-data methods. MILC (AIC) was slightly more stable than MILC (Large), and MICE (LOG) is slightly more stable than MICE (PMM). This was expected because MILC (AIC) and MICE (LOG) are more parsimonious than MILC (Large) and MICE (PMM), respectively. The expected result that MILL would be less stable than MLID was not demonstrated. As expected, sample size had a positive effect on stability. For small samples ($N = 200$), the stability could be considered low, resulting in low power. For example,

Table 2. Stability of the estimates of three logistic regression coefficients for six incomplete-data methods, three different levels of strength of three-variable associations ($\beta_{23} = 0$, $\beta_{23} = -.8$, $\beta_{23} = -2$), two sample sizes (200, 1000), and 40% missingness

			Incomplete-data method					
N	β_{23}	RC	MLID	MILL	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
200	0	$\beta_2 = 1$	.862	.845	.858	.600	.951	.738
		$\beta_4 = 1$	.500	.496	.502	.445	.518	.501
		$\beta_{23} = 0$	1.19	1.17	1.17	.840	1.23	.729
	−.8	$\beta_2 = 1.4$	.938	.931	.924	.679	1.01	.780
		$\beta_4 = 1$	.505	.506	.510	.448	.533	.508
		$\beta_{23} = -.8$	1.21	1.22	1.20	.862	1.26	.737
	−2	$\beta_2 = 2$	.916	.956	.948	.748	1.02	.806
		$\beta_4 = 1$	.494	.510	.515	.449	.537	.501
		$\beta_{23} = -2$	1.24	1.25	1.25	.917	1.29	.771
1000	0	$\beta_2 = 1$	.344	.373	.370	.264	.380	.301
		$\beta_4 = 1$	.203	.206	.206	.188	.206	.205
		$\beta_{23} = 0$	.479	.515	.509	.362	.522	.306
	−.8	$\beta_2 = 1.4$	.365	.389	.384	.291	.392	.313
		$\beta_4 = 1$	.205	.207	.208	.188	.207	.204
		$\beta_{23} = -.8$	.470	.507	.494	.376	.501	.310
	−2	$\beta_2 = 2$	.377	.407	.404	.342	.402	.306
		$\beta_4 = 1$	.200	.205	.205	.185	.205	.197
		$\beta_{23} = -2$	.482	.526	.517	.451	.514	.298

Note: N: sample size; β_{23}: strength of three-variable association; RC: regression coefficient.

the population value of β_4 was equal to 1, but in case $\text{sd}(\hat{\beta}_4) = .501$ (MILL, medium three-variable association), which is even one of the smaller standard deviations we found, one may expect to find estimates of β_4 between .02 and 1.98 (95% confidence interval). For large samples ($N = 1000$), the stability is much better. Percentage of missingness also had a negative effect on the stability. This can be expected because a larger percentage of missingness in fact means a reduction of the sample size.

3.2.3 Bias of the standard errors

Table 3 shows the most important results for bias of the standard errors for 20% missingness. Bias of the standard errors was smaller for $N = 1000$ than for $N = 200$. Bias of the standard errors was largest for the parameters associated with the three-variable association (β_2 and β_{23}) For $N = 200$, MILL and MILC (Large) had the smallest bias, whereas MILC (AIC) and MICE (LOG) overestimated the standard errors and MLID and MICE (PMM) underestimated the standard errors. For $N = 1000$, MILL, MILC (AIC), MILC (Large), and MICE (PMM) had the smallest bias, whereas MLID underestimated and MICE (LOG) overestimated the standard errors. This renders MILC (Large) as the most favorable incomplete-data method with respect to bias in standard errors.

Table 3. Bias in the standard errors of the estimates of three logistic regression coefficients for six incomplete-data methods, three different levels of strength of three-variable associations ($\beta_{23} = 0$, $\beta_{23} = -.8$, $\beta_{23} = -2$), two sample sizes (200, 1000), and 40% missingness.[a]

			Incomplete-data method					
N	β_{23}	RC	MLID	MILL	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
200	0	$\beta_2 = 1$	−.040	−.044	−.044	**.151**	**−.202**	.072
		$\beta_4 = 1$	−.025	−.021	−.022	.015	−.052	−.021
		$\beta_{23} = 0$	−.092	**−.108**	−.077	**.174**	**−.219**	**.296**
	−.8	$\beta_2 = 1.4$	−.086	−.077	−.064	**.116**	**−.214**	.073
		$\beta_4 = 1$	−.027	−.017	−.018	.022	−.054	−.018
		$\beta_{23} = -.8$	−.092	−.077	−.062	**.194**	**−.209**	**.315**
	−2	$\beta_2 = 2$	−.032	−.065	−.045	**.080**	**−.199**	.065
		$\beta_4 = 1$	−.010	−.015	−.017	.024	−.052	−.012
		$\beta_{23} = -2$	−.074	−.064	−.054	**.164**	**−.193**	**.292**
1000	0	$\beta_2 = 1$	−.001	−.015	−.011	**.065**	**−.060**	.038
		$\beta_4 = 1$	−.006	−.005	−.005	.008	−.010	−.005
		$\beta_{23} = 0$	−.015	−.030	−.025	**.088**	**−.086**	**.130**
	−.8	$\beta_2 = 1.4$	−.017	−.023	−.021	.041	−.067	.029
		$\beta_4 = 1$	−.007	−.006	−.007	.009	−.011	−.004
		$\beta_{23} = -.8$	−.005	−.015	−.005	**.070**	**−.059**	**.126**
	−2	$\beta_2 = 2$	−.012	.040	−.036	**.080**	**−.074**	.032
		$\beta_4 = 1$	.002	−.002	−.000	.024	−.006	.001
		$\beta_{23} = -2$	.002	.028	−.028	**.164**	**−.068**	**.133**

Note: N = sample size; β_{23} = strength of three-variable association; RC = regression coefficient.
[a]Remarkable bias is printed in boldface.

4 Study 2: bias, stability, and bias in standard errors produced by MILC, MICE and complete-case analysis for a larger number of trichotomous variables

In Study 2, we compared incomplete-data methods MILC (AIC), MILC (Large, $K = 33$), MICE (PMM), and MICE (LOG) to complete-case analysis, on bias of the parameter estimates, stability of the parameter estimates, and bias of the standard errors. Benchmarks MLID and MILL could no longer be used because the number of variables (11) was too large. The main question for Study 2 was whether MILC and MICE would also work for polytomous categorical data and for large numbers of possible response patterns. In Study 1, the number of possible response patterns was $2^6 = 64$, whereas in Study 2, the number of possible response patterns was increased to $3^{11} = 177{,}147$. The main objective for the design of Study 2 is that the associations among the variables need to be complex, to test whether the incomplete-data methods can pick up the associations correctly.

4.1 Method

4.1.1 General set-up

In Study 2, the population model from which the complete data sets were sampled contained eight trichotomous predictor variables ($Y_1, \ldots, Y_8$) and three trichotomous outcome variables (Y_9, Y_{10}, and Y_{11}). The categories were coded -1, 0, and 1. The associations among the 11 are described by a path model for categorical data[46] containing one-, two-, and three-way associations (see Figure 1 for a graphical representation and Appendix A for the chosen parameter values).

Variables Y_1, Y_3, Y_4, and Y_{11} had missing values; the other variables were completely observed. The missingness mechanism was MAR, and rather complex. For Y_1 and Y_3, the missingness depended on Y_2 and Y_9. Let R indicate whether (score 1) or not (score 0) a score is observed. Both for Y_1 and Y_3, the logit of R was $\text{logit}(R) = -5.06 + -2 \cdot Y_9 + 3 \cdot Y_2$, resulting in approximately 20% missing values for each variable. Similarly, for Y_4 and Y_{11}, the missingness depended on Y_7 and Y_9. Both for Y_4 and Y_{11}, $\text{logit}(R) = -5.50 + 3 \cdot Y_9 - 1.5 \cdot Y_7$, also resulting in approximately 20% missing values for each variable. This procedure kept the total percentage of missingness constant at $4/11 \cdot 20\% + 7/11 \cdot 0\% = 7.27\%$.

For each incomplete data set, the multiple imputation methods created $m = 5$ completed data sets. For complete-case analysis, a complete data set was obtained by simply deleting every row that contained at least one missing value.

The substantive model was an adjacent category ordinal logit model[35] for outcome variable Y_9 containing Y_1, Y_3, Y_4, and Y_{11} as predictors. The logit equation has the form

$$\text{logit}(Y_9 = j \,|\, Y_9 = j-1 \text{ or } Y_9 = j) = \beta_{0j} + \beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \beta_{12} Y_1 Y_2,$$

for $j = 1, 2$. Note that the substantive model is part of the population model (Figure 1) and includes the main effects of the predictors of Y_9, and the interaction effect of Y_1 and Y_2. The latter implies a three-variable association among Y_1, Y_2, and Y_9.

Three software packages were used for data generation, incomplete-data handling, and estimating the substantive model. Complete and incomplete data were generated by LEM.[20] The imputation phase of MILC (Large) and MICE (PMM) was performed using Latent GOLD and the R package MICE, respectively. Latent GOLD was used to estimate the substantive model for by MILC (Large) and MICE (PMM), using the completed datasets, and for MLID and complete-case analysis.

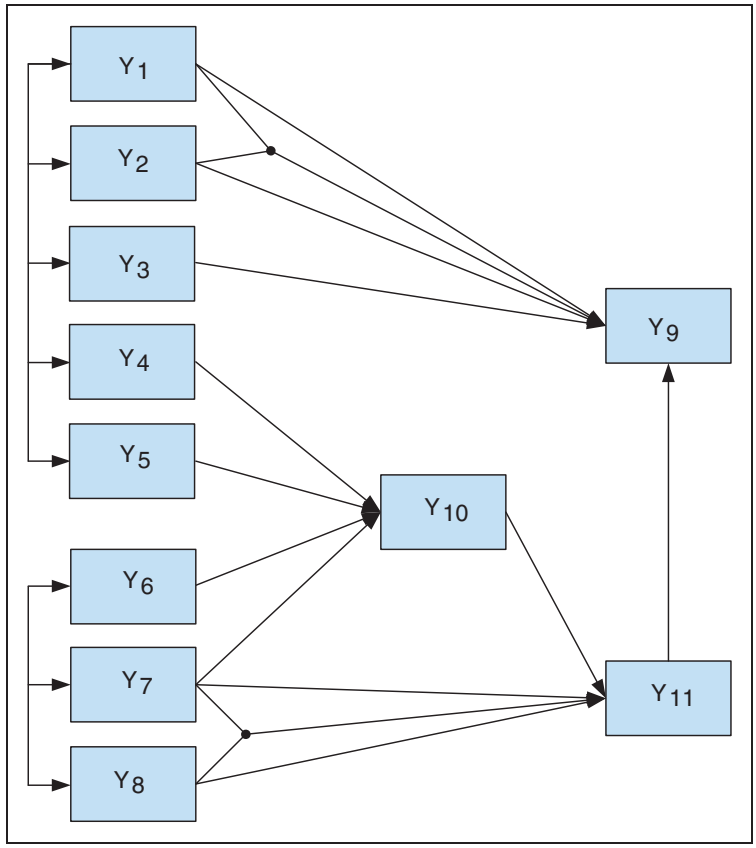

Figure 1. Population model of the second simulation study. The model contains 11 trichotomous variables: Y_1 through Y_8 are predictor variables and Y_9 through Y_{11} are outcome variables.

4.1.2 Design

We only varied sample size and incomplete-data method. Sample size had two levels: medium ($N = 500$) and large ($N = 1000$); incomplete-data method had five levels: MILC (AIC), MILC (Large), MICE (LOG), MICE (PMM), and complete-case analysis. This yields a 5×2 design. The outcome variables were equivalent to those in Study 1 (bias, stability, and bias of standard errors).

4.2 Results

4.2.1 Bias

Table 4 shows the bias for β_2, the main effect of a predictor that is also involved in the interaction effect; β_3, the main effect of a predictor not involved in the interaction effect; and β_{12}, the interaction effect itself. The most important result is that MICE (PMM) and MICE (LOG) produced relatively large bias in the estimates of β_3 and β_{12}, suggesting that the imputation models of MICE (LOG) and MICE (PMM) do not correctly pick up the three-way association in the data. Furthermore, complete-case analysis produced very large bias in the estimates of β_2 and β_{12}, confirming that

Table 4. Bias in the estimates of three logistic regression coefficients for five incomplete-data methods, two sample sizes (500, 1000), and 20% missingness on four variables.[a]

		Incomplete-data method				
N	RC	CC	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
500	$\beta_2 = -.45$	**.504**	−.033	−.028	−.035	−.036
	$\beta_3 = .5$	.017	.001	−.002	**.054**	**.051**
	$\beta_{12} = .45$	**−.110**	−.068	−.066	**−.114**	**−.113**
1000	$\beta_2 = -.45$	**.501**	−.027	−.025	−.033	−.035
	$\beta_3 = .5$	.019	.001	−.003	**.053**	**.046**
	$\beta_{12} = .45$	**−.113**	−.061	−.061	**−.116**	**−.114**

Note: *N*: sample size; RC: regression coefficient; CC: complete-case analysis.
[a] Remarkable bias is printed in boldface.

Table 5. Stability of the estimates of three logistic regression coefficients for five incomplete-data methods, two sample sizes (500, 1000), and 20% missingness on four variables

		Incomplete-data method				
N	RC	CC	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
500	$\beta_2 = -.45$	.101	.290	.288	.293	.292
	$\beta_3 = .5$	.215	.218	.215	.240	.240
	$\beta_{12} = .45$	.146	.159	.157	.127	.130
1000	$\beta_2 = -.45$	.069	.272	.271	.275	.275
	$\beta_3 = .5$	.224	.222	.220	.247	.246
	$\beta_{12} = .45$	.134	.162	.161	.134	.137

Note: *N*: sample size; RC: regression coefficient; CC: complete-case analysis.

complete-case analysis leads to biased results when the data are MAR. MILC (AIC) and MILC (Large) had a similar performance in terms of bias.

4.2.2 Stability

Table 5 shows the stability of β_2, β_3, and β_{12}. The most important result is that (almost) unbiased parameter estimates (see Table 4) showed similar stability across methods, whereas biased parameter estimates tended to be either more stable or more unstable. This effect was clearer for $N = 500$ than for $N = 1000$. For example, for the estimate of β_2, MILC (Large), MILC (AIC), MICE (PMM), and MICE (LOG) show similar bias and similar stability of parameter estimates. However, complete-case analysis overestimated β_2 and this estimate was too stable, whereas MICE (PMM) and MICE (LOG) overestimated β_3 and this estimate was too unstable.

4.2.3 Bias of the standard errors

Table 6 shows the bias of the standard errors of β_2, β_3, and β_{12}. Bias of the standard errors was smaller for $N = 1000$ than for $N = 500$. The multiple imputation methods yielded similar upward bias in their standard errors, whereas complete-cases analysis tended to yield a larger overestimation of the standard errors. Bias was largest for the standard error of the interaction effect (β_{12}).

Table 6. Bias in the standard errors of the estimates of three logistic regression coefficients for five incomplete-data methods, two sample sizes (500, 1000), and 20% missingness on four variables[a]

		Incomplete-data method				
N	RC	CC	MILC (large)	MILC (AIC)	MICE (PMM)	MICE (LOG)
500	$\beta_2 = -.45$	**.126**	.080	.080	.082	.081
	$\beta_3 = .5$	**.120**	.081	.083	.088	.084
	$\beta_{12} = .45$	**.156**	.100	.102	.106	.102
1000	$\beta_2 = -.45$	**.089**	.058	.058	.058	.057
	$\beta_3 = .5$	**.086**	.061	.062	.062	.059
	$\beta_{12} = .45$	**.111**	.076	.077	.075	.072

Note: *N*: sample size; RC: regression coefficient; CC: complete-case analysis.
[a]Remarkable bias is printed in boldface.

Table 7. Information on the variables of the Projective Services Project for older persons.[47]

Variable	Levels	Code
Mental status	Poor, good	Y_1
Physical status	Poor, good	Y_2
Age	less than 75, over 75	Y_3
Group membership	experiment, control	Y_4
Sex	male, female	Y_5
Survival status	Deceased, survived	Y_6

5 Real-data example

We applied the most promising variants of MILC and MICE (i.e., MILC (Large, $K = 12$) and MICE (PMM)), complete-case analysis, and MLID to data from the Investigators of Projective Services Project for Older Persons,[47] which have been discussed and analyzed earlier by Fuchs[48] to illustrate the MLID approach. The data set contains the scores of 164 patients on six dichotomous variables (Table 7). One patient had a missing value on the physical status, 33 had a missing value for mental status, and 29 respondents had a missing value on both physical and mental status. The question of interest is whether the unexpected negative association between treatment and survival disappears when controlling for age, gender, physical status, and mental status.

The substantive model predicts survival by the main effects of all variables, plus the interaction effect of mental status, Y_1, and physical status, Y_2. We defined the following regression model,

$$\text{logit}(Y_6) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \beta_3 Y_3 + \beta_4 Y_4 + \beta_5 Y_5 + \beta_{12} Y_1 Y_2. \quad (7)$$

Contrary to Fuchs, we chose to include the interaction between mental status and physical status because we were interested in whether the imputation methods yielded similar results to MLID in a model containing a higher-order association. Once the data had been imputed using MILC (Large)

Table 8. Estimated logistic regression coefficients using MLID, MILC (Large), and MICE (PMM)[a]

RC	Incomplete-data method							
	MLID		Complete-case		MILC (Large)		MICE (PMM)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
β_1	3.175	2.188	2.844	2.070	3.777	1.984	**2.271**	1.815
β_2	2.614	2.240	**1.816**	2.180	3.162	2.102	**1.439**	1.979
β_3	−1.417	.431	−1.568	.526	−1.380	.422	−1.426	.434
β_4	.459	.394	**.176**	.481	.496	.392	.475	.384
β_5	−.506	.417	**−.281**	.525	−.420	.437	−.583	.410
β_{12}	−1.017	1.269	**−.629**	1.217	−1.283	1.168	**−.382**	1.090

Note: RC: regression coefficient; Est: estimate; SE: standard error.
[a]Remarkable bias is printed in boldface

and MICE (PMM), the substantive model defined in equation (7) was estimated. We also estimated the logistic regression model using MLID (as a benchmark) and complete-case analysis. Schafer's notion on the number of variables is of no concern in this analysis because the substantive model and the imputation model are identical; both models include all available variables. Therefore, the performance of MILC (Large), MICE (PMM), and complete-case analysis was assessed by comparing them to MLID (Table 8).

For all incomplete-data methods, only age (β_3, negative effect) had a significant effect on survival status. The fact that all other effects were not statistically significant may be due to the small sample size. Nevertheless, it remains interesting to compare the parameter estimates across incomplete-data methods. Table 8 shows that the estimates yielded by MILC (Large) were very similar to MLID, for all parameters. MICE (PMM) produced estimates of β_3, β_4 and β_5 that were very similar to MLID, but yielded relatively large differences for parameters β_1, β_2, and β_{12}. Complete-case analysis produced rather large differences for the estimates of $\beta_2, \beta_4, \beta_5$ and β_{12}. MLID, MILC (Large), and MICE (PMM) did not have large differences in the estimated standard errors. However, complete-case analysis yielded relatively large standard errors for parameters $\beta_3, \beta_4,$ and β_5, compared to MLID.

6 Discussion

The aim of this paper was to investigate which incomplete-data method for categorical data should be recommended to practitioners. We assessed the performance of MILC and MICE with regard to three criteria, relative to MLID, MILL, and complete-case analysis. Based on the theoretical discussion, Study 1, and Study 2, MILC (Large) appears to be the incomplete-data method that meets the three criteria to the greatest extent. The other incomplete-data methods have one or more features that make them suboptimal. MILL cannot handle more than a few variables, MLID does not allow for the use of small substantive model as it can affect the MAR assumption, and may yield biased parameter estimates for a complex association in case of a small sample size. While in Study 1 MICE (PMM) performed rather well, Study 2 showed that MICE (PMM) may yield biased parameter estimates when the number of possible data pattern is large, especially when the sample size is small. MILC (AIC) and MICE (LOG) may fail to capture higher-order associations in the data, which yields parameter estimates with an unacceptably high bias.

In Study 2 it was demonstrated that complete-case analysis yields very large bias in the parameter estimates and a loss of power due to inflated standard errors. The findings in the real-data example were consistent with these results.

A remaining issue with MILC is that there is not yet a guideline indicating the minimum number of required latent classes. The simulation study showed that over fit does not seem to be a problem, which was also argued by Vermunt et al.,[10] so one can always resort to estimating a latent class model with many classes. However, having a minimum of required latent classes would greatly facilitate the use of MILC because estimating latent class models with 40, 50, or 60 latent classes can be very time consuming. We showed that in case of a small table AIC is not a good criterion because the number of classes is too low; for a large table MILC (AIC) yielded good results. A heuristic rule may be to use as many classes as there are categories in the data. For example, for a data set consisting of 10 variables with three response categories and 5 dichotomous variables, the number of latent classes would be $10 \times 3 + 5 \times 2 = 40$ latent classes. Whether this heuristic rule is reasonable should be investigated in future research.

An additional comment should be made for MICE (LOG), as it may have been presented too negatively. The problem of MICE (LOG) is that in the default setting, interaction effects are not included in the conditional models. As a result, the imputation model may be too parsimonious yielding biased parameter estimates. Further research would be required to investigate whether this method is able to produce unbiased results if the conditional model included higher-order interactions.

Lastly, we note that each incomplete-data method had to be applied using a different software package, as there is no package available that applies all of the methods. Further research is warranted to investigate the potential differences between implementations of MILC and MICE across software packages.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Daniël W van der Palm is supported by a grant from the Netherlands Organisation for Scientific Research (NWO 400-08-234).

References

1. Kurian A, Gallagher S, Cheeyandira A, et al. Predictors of in-hospital length of stay after laparoscopic ventral hernia repair: results of multivariate logistic regression analysis. *Surg Endosc* 2010; **24**: 2789–2792.
2. Zhu B, Walter SD, Rosenbaum PL, et al. Structural equation and log-linear modeling: a comparison of methods in the analysis of a study on caregivers' health. *BMC Med Res Methodol* 2006; **6**: 49.
3. Luciano JV, Ayuso-Mateos JL, Aguado J, et al. The 12-item World Health Organization disability assessment schedule II (WHO-DAS II): a nonparametric item response analysis. *BMC Med Res Methodol* 2010; **10**: 45.
4. Klebanoff MA and Cole SR. Use of multiple imputation in the epidemiologic literature. *AM J Epidemiol* 2008; **168**: 355–357.
5. Little RJA and Rubin DB. *Statistical analysis with missing data*, 2nd ed. New York: Wiley, 2002, pp.266–291.
6. Schafer JL. *Analysis of incomplete multivariate data*. London: Chapman & Hall, 1997.
7. Schafer JL and Graham JW. Missing data: our view of the state of the art. *Psychol Methods* 2002; **7**: 147–177.
8. *SPSS Inc. SPSS 19 for Windows. Somers*. NY: IBM, 2011.
9. *SAS Inc. SAS for Windows. Cary*. NC: SAS, 2011.
10. Vermunt JK, Van Ginkel JR, Van der Ark LA, et al. Multiple imputation of incomplete categorical data using latent class analysis. *Sociol Methodol* 2008; **38**: 369–397.
11. Gebregziabher M and DeSantis SM. Latent class based multiple imputation approach for missing categorical data. *J Stat Plan Inference* 2010; **140**: 3252–3262.

12. Van Buuren S and Groothuis-Oudshoorn K. Multivariate imputation by chained equations in R. *J Stat Soft* 2011; **45**.
13. Van Buuren S, Brand PL, Groothuis-Oudshoorn K, et al. Fully conditional specification in multivariate imputation. *J Stat Comput Sim* 2006; **76**: 1049–1064.
14. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 2007; **16**: 219–242.
15. Dempster AP, Laird NM and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol* 1977; **39**: 1–38.
16. Allison PD. *Missing data*. Thousand Oaks, CA: Sage, 2001.
17. Arbuckle JL. Full information estimation in the presence of incomplete data. In: Marcoulides GA and Schumacker SE (eds) *Advanced Structural Equation Modeling*. Mahwah, NJ: Erlbaum, 1996, pp.243–277.
18. Ezzati-Rice TM, Johnson W, Khare M, et al. A simulation study to evaluate the performance of model-based multiple imputations in NCHS Health Examination Surveys. In: *Proceedings of the Bureau of the Census Eleventh Annual Research Conference*, 1995, pp.257–266.
19. Schafer JL, Ezzati-Rice TM, Johnson W, et al. The NHANES III multiple imputation project. In: *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1998, pp.28–37.
20. Vermunt JK. *LEM: A general program for the analysis of categorical data*. The Netherlands: Tilburg: Department of Methodology and Statistics, Tilburg University, 1997.
21. Muthén LK and Muthén BO. *Mplus 6.1*. Los Angeles: Muthén & Muthén, 2010.
22. White IR, Royston P and Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2010; **30**: 377–399.
23. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: Wiley, 1987.
24. Efron B and Tibshirani R. *An introduction to the bootstrap*. London: Chapman & Hall, 1993.
25. Vermunt JK and Magidson J. *LG-syntax user's guide: Manual for Latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations Inc, 2008.
26. Lazarsfeld PF. The logical and mathematical foundation of latent structure analysis. In: Stouffer SA, Guttman L, Suchman EA, et al (eds) *Studies in social psychology in World War II. Vol. IV: Measurement and prediction*. Princeton: Princeton University Press, 1950, pp.361–412.
27. Lazarsfeld PF. The interpretation and mathematical foundation of latent structure analysis. In: Stouffer SA, Guttman L, Suchman EA, et al (eds) *Measurement and prediction*. Princeton: Princeton University Press, 1950, pp.413–472.
28. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974; **64**: 215–231.
29. Vermunt JK and Magidson J. Latent class analysis. In: Lewis-Beck MS, Bryman AE and Liao TF (eds) *The Sage Encyclopedia of Social Science Research Methods*. Thousand Oaks: Sage, 2004, pp.549–553.
30. McLachlan GJ and Peel D. *Finite mixture models*. New York: Wiley, 2000.
31. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 1987; **52**: 345–370.
32. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–464.
33. Bozdogan H. Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse-fisher information matrix. In: Opitz O, Lausen B and Klar R (eds) *Information and classification, concepts, methods and applications*. Berlin: Springer, 1993, pp.40–54.
34. Drechsler J and Rassler S. Does convergence really matter? In: Shalabh CH (ed.) *Recent advances in linear models and related areas. Essays in honour of Helge Toutenburg*. Berlin: Springer, 2008, pp.341–355.
35. Agresti A. *Categorical data analysis*. New York: Wiley, 1990.
36. *StataCorp. Stata statistical software: Release 12. College Station*. TX: StataCorp, 2011..
37. Royston P. Multiple imputation of missing values: Further update of ICE, with an emphasis on categorical variables. *Stata J* 2009; **9**: 466–477.
38. Ibrahim JG, Chen MH, Lipsitz SR, et al. Missing data methods in generalized linear models: a comparative review. *J Am Stat Assoc* 2005; **100**: 332–346.
39. Kang JDY and Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci* 2007; **22**: 523–539.
40. Vansteelandt S, Carpenter J and Kenward MG. Analysis of incomplete data using inverse probability weighting and double robust estimators. *Eur J Res Methods Behav Soc Sci* 2010; **6**: 37–48.
41. Graham JW and Schafer JL. On the performance of multiple imputation for multivariate data wit small sample size. In: Hoyle R (ed.) *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage, 1999, pp.1–29.
42. Benaards CA, Belin TR and Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med* 2007; **26**: 1368–1382.
43. Allison PD. Imputation of categorical variables with PROC MI. *SUGI 30 Proceedings* 2005; **113–30**: 1–14.
44. Horton NJ, Lipsitz SP and Parzen MA. Potential for bias when rounding in multiple imputation. *Am Stat* 2003; **57**: 229–232.
45. Neyman J and Pearson ES. On the problem of most efficient tests of statistical hypotheses. *Phil Trans R Soc London Series A* 1933; **231**: 289–337.
46. Goodman LA. The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika* 1973; **60**: 179–192.
47. Blenkner M, Bloom M and Weber R. Final report: Protective services for older people. Cleveland, OH: Benjamin Rose Institute, 1974.
48. Fuchs C. Maximum likelihood estimation and model selection in contingency tables with missing data. *J Am Stat Assoc* 1982; **77**: 270–278.

Appendix A

Tables A1, A2, and A3 show the parameter values describing the population model in Study 2 (Figure 1). All variables had three nominal response categories. Because dummy coding was used, the effect of the first category was zero (not displayed). For all two-way and three-way interactions only a single value is shown because the associations were defined to be ordinal (linear-by-linear).

Table A1. Loglinear parameters describing $P(Y_1,Y_2,Y_3,Y_4,Y_5)$

$Y_1 = (.15,.0)$	$Y_1Y_2 = .3$	$Y_2Y_4 = .55$	$Y_1Y_4Y_5 = -.35$
$Y_2 = (.25,.05)$	$Y_1Y_3 = -.4$	$Y_2Y_5 = -.36$	$Y_2Y_4Y_5 = -.25$
$Y_3 = (-.15, -.15)$	$Y_1Y_4 = -.2$	$Y_3Y_4 = -.15$	$Y_1Y_2Y_3 = .55$
$Y_4 = (.05,.25)$	$Y_1Y_5 = .5$	$Y_3Y_5 = -.05$	
$Y_5 = (-.25, -.05)$	$Y_2Y_3 = -.3$	$Y_4Y_5 = .3$	

Table A2. Loglinear parameters describing $P(Y_6,Y_7,Y_8)$

$Y_6 = (.15,.75)$	$Y_6Y_7 = .32$
$Y_7 = (.25,1.25)$	$Y_6Y_8 = -.4$
$Y_8 = (.1,.5)$	$Y_7Y_8 = .24$
	$Y_6Y_7Y_8 = .4$

Table A3. Logistic regression parameters

$Y_9\|Y_1 = -.3$	$Y_{10}\|Y_4 = .22$	$Y_{11}\|Y_{10} = -.15$
$Y_9\|Y_2 = -.45$	$Y_{10}\|Y_5 = .32$	$Y_{11}\|Y_6 = -.3$
$Y_9\|Y_3 = .5$	$Y_{10}\|Y_6 = .42$	$Y_{11}\|Y_7 = .35$
$Y_9\|Y_1Y_2 = .45$	$Y_{10}\|Y_7 = -.38$	$Y_{11}\|Y_8 = .1$
$Y_9\|Y_{11} = .35$	$Y_{10}\|Y_8 = .34$	$Y_{11}\|Y_6Y_7 = .4$
	$Y_{10}\|Y_6Y_7 = -.14$	

Table A1 shows the log linear parameters describing $P(Y_1, Y_2, Y_3, Y_4, Y_5)$, Table A2 shows the log linear parameters describing $P(Y_6, Y_7, Y_8)$, and Table A3 shows the logistic regression parameters relating predictor variables $Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7$, and Y_8 to outcome variables Y_9, Y_{10}, and Y_{11}.