American Journal of Epidemiology
© The Author 2010. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

Vol. 171, No. 5
DOI: 10.1093/aje/kwp425
Advance Access publication:
January 27, 2010

Practice of Epidemiology

Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation

Katherine J. Lee* and John B. Carlin

* Correspondence to Dr. Katherine J. Lee, Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, Royal Children's Hospital, Flemington Road, Parkville, Victoria 3052, Australia (e-mail: katherine.lee@mcri.edu.au).

Initially submitted June 22, 2009; accepted for publication November 30, 2009.

Statistical analysis in epidemiologic studies is often hindered by missing data, and multiple imputation is increasingly being used to handle this problem. In a simulation study, the authors compared 2 methods for imputation that are widely available in standard software: fully conditional specification (FCS) or ''chained equations'' and multivariate normal imputation (MVNI). The authors created data sets of 1,000 observations to simulate a cohort study, and missing data were induced under 3 missing-data mechanisms. Imputations were performed using FCS (Royston's ''ice'') and MVNI (Schafer's NORM) in Stata (Stata Corporation, College Station, Texas), with transformations or prediction matching being used to manage nonnormality in the continuous variables. Inferences for a set of regression parameters were compared between these approaches and a complete-case analysis. As expected, both FCS and MVNI were generally less biased than complete-case analysis, and both produced similar results despite the presence of binary and ordinal variables that clearly did not follow a normal distribution. Ignoring skewness in a continuous covariate led to large biases and poor coverage for the corresponding regression parameter under both approaches, although inferences for other parameters were largely unaffected. These results provide reassurance that similar results can be expected from FCS and MVNI in a standard regression analysis involving variously scaled variables.

data interpretation; statistical; epidemiologic methods; imputation; incomplete data; missing data; simulations

Abbreviations: FCS, fully conditional specification; MVNI, multivariate normal imputation.

Multiple imputation has become increasingly popular for handling missing data in epidemiologic analysis (1, 2). Initially, statistical models are used to obtain plausible substitutes for missing values, with the imputation process being repeated several times to allow for the uncertainty in the missing values. Analytic results are then obtained by combining the results of standard complete-data analyses across the multiple completed data sets in an appropriate manner (3, 4). Implementation of the second stage of this process is straightforward, with software tools available to facilitate the awkward procedure of analyzing several copies of the original data set (5–7). However, the first stage is relatively complex, and although there is a range of software tools that perform imputation (8), they each have limitations.

Missing data commonly occur in a general pattern (nonmonotone missingness), and there are currently 2 widely available methods of model-based imputation that are used with such data sets: multiple imputation based on the multivariate normal distribution (MVNI), originally implemented by Schafer (4), and the method based on fully conditional specification (FCS), also known as ''chained equations'' or ''regression switching,'' implemented independently by van Buuren et al. (9) and Raghunathan et al. (10, 11) (see also Royston (12, 13)). These 2 approaches are based on quite different theoretical assumptions and involve very different computational methods. However, it is unclear how important these differences are in practice, since both have features that in principle make their application invalid or potentially unreliable for particular problems. Although a number of comparisons between them have been published (14–16), these have focused on aspects that may not be of primary importance in many epidemiologic applications, and each approach has not always been implemented in the best possible way.

In this paper, we compare MVNI and FCS approaches to multiple imputation in the context of estimating coefficients in a linear regression model (including adjustment for potential confounding), along the lines of a typical epidemiologic analysis. We were particularly interested in exploring whether the stricter model assumptions of MVNI make it less robust than the more flexible FCS. Simulated data were created from a realistically complex model, with missing data being subsequently imposed, and the results of the regression analyses under the alternative imputation methods were compared. We also examined the effects of different approaches to dealing with nonnormality in a continuous variable in these imputation methods.

BACKGROUND: OVERVIEW OF APPROACHES TO IMPUTATION METHODS

Multivariate normal imputation

The first widely available general-purpose imputation algorithm was published by Schafer (4) and made available in his stand-alone ‘‘NORM’’ package (17). This method is now available within a number of statistical packages, including SAS (7) and Stata (18). MVNI assumes that all variables in the imputation model jointly follow a multivariate normal distribution. Implementation uses a Bayesian approach (with a Markov chain Monte Carlo algorithm) to obtain imputed values from the estimated multivariate normal distribution, allowing appropriately for uncertainty in the estimated model parameters, as required for ‘‘proper’’ imputation (3). Clearly, the assumption of multivariate normality is often not plausible, especially in the presence of binary and categorical variables. However, Schafer (4) suggested that inference from MVNI may often be reasonable even if multivariate normality does not hold, and MVNI has been widely applied in contexts where data are clearly not multivariate normal (19, 20).

Fully conditional specification

The fully conditional approach to imputation is a more flexible method that does not rely on the assumption of multivariate normality (9, 11). Conditional distributions (regression models) are specified for each variable with missing values, conditional on all of the other variables in the imputation model. Imputations are generated by estimating each conditional distribution in turn, using observed cases for the variable being considered and imputed values for the other variables at that iteration and imputing missing values (again allowing for uncertainty in model parameters). The approach is appealing, since it does not restrict the conditional distributions to being normal, so that univariate regression models can be tailored appropriately—for example, using logistic regression for binary variables and ordered logistic regression for ordinal variables. However, it is possible for some of the conditional distributions to be incompatible with each other, potentially leading to unsound imputations (14). The FCS approach is also available in a number of statistical packages, including Stata (13), SAS (11), and R (21), and is increasingly appearing in practice (22, 23).

Nonnormal continuous variables and prediction matching

In both approaches, continuous variables with clearly nonnormal (skewed) distributions are unlikely to be handled adequately without special treatment, since multivariate normality implies a normal marginal distribution for each variable and standard FCS draws imputed values for a continuous variable using a normal linear regression on the other variables specified. An alternative for FCS is to impute values using prediction matching (12, 24), under which the missing value is replaced by the nonmissing value for the case whose predictive mean is closest to that of the case with the missing value. Within MVNI, a natural approach to skewness is to utilize (univariate) normalizing transformations; these can also be applied when using FCS.

METHODS

Simulated data set and target analysis

Simulated complete data sets were obtained by random sampling from a synthetic ‘‘population’’ of 971,327 girls created to resemble data from the US National Longitudinal Study of Adolescent Health (25). The National Longitudinal Study of Adolescent Health is a study of 6,000 US girls in grades 7–11 who were recruited during the 1994–1995 school year. Full details on this synthetic population are presented elsewhere (25). Variables in the data set resembled items from successive annual waves of the original study, with the primary outcome being a measure of emotional distress at wave II (continuous variable with range 0–3; higher scores represent greater distress). The primary covariate was a binary indicator for whether or not the girl had dieted in the previous 7 days at wave I, with a further 13 covariates including race (categorized as black, nonblack Hispanic, and other) and grade (ordinal variable with range 7–11). The variables were synthesized sequentially, beginning with drawing 1 million observations from the 3 × 5 contingency table of race and grade and then adding other variables one at a time using predictive simulation from regression models based on the original data set. At each step, the model conditioned on the previously generated values, incorporating them into complex regressions that included nonlinear relations and numerous interactions, with the aim of creating a ‘‘population’’ that had realistic complexity. To simplify the handling of logarithmic transformations in our study, we excluded cases with emotional distress scores of 0 at either wave I or wave II (2.9% of the original 1 million).

Data sets for this study were created by drawing random samples of 1,000 from the synthetic population. We focused on estimating the coefficient of the primary covariate, the dieting indicator, in a regression model for emotional distress at wave II. Other covariates included in the model were baseline emotional distress (continuous variable with range 0–3), race (2 indicator variables), grade, self-rated overall health (ordinal variable with range 1–5), and self-rated physical fitness (ordinal variable with range 1–5). Since the distress outcome was strongly positively skewed, we followed the standard approach of carrying out the analysis on the (natural) log scale, applying the same transformation to the outcome (wave II) and baseline (wave I) distress values. Thus, the analysis of interest was the linear regression model:

$$\begin{aligned} ldistW2 = \alpha + \beta_1 diet + \beta_2 \mathrm{l}distW1 + \beta_3 race_1 \\ + \beta_4 race_2 + \beta_5 grade + \beta_6 health + \beta_7 fitness, \end{aligned} \tag{1}$$

where *ldistW1* and *ldistW2* represent the log-transformed distress measures at waves I and II, respectively, *diet* is the dieting indicator, $race_1$ indicates black race, $race_2$ indicates nonblack Hispanic race, *grade* is an integer in the range 7–11, and *health* and *fitness* are the ordinal measures of health and physical fitness, the latter 3 fitted as linear effects.

The "true values" of the regression coefficients were defined to be the least-squares estimates obtained in the full synthetic population. In the original population, the (adjusted) effect of diet on emotional distress was very small ($\beta_1 = -0.028$). In order to ensure that our findings were not limited by being restricted to a null hypothesis setting, for our main results we artificially inflated the diet effect to 10% on the log scale ($\beta_1 = -0.101$), so that it was borderline statistically significant with the chosen sample size. As a sensitivity analysis, we also considered the original simulated data set with the very small diet effect.

Missing-data models

We considered 3 missing-data models, with a progressive increase in both the number of variables with missing data and the number of incomplete observations:

Model 1: Missing data on emotional distress at wave I only.

Model 2: As for model 1, with missing data also on health and physical fitness.

Model 3: As for model 2, with missing data also on diet.

For missing-data models 2 and 3, we assumed that self-rated health and physical fitness were questionnaires the students completed at the same time, so that data for both would be either missing or nonmissing. Missingness in diet and missingness in distress at wave I were assumed to occur independently of each other and of health and fitness. In each case, values were set to missing with a probability determined by a logistic regression model dependent on the outcome, emotional distress at wave II, race and grade (all fully observed), and diet:

$$\begin{aligned} \operatorname{logit} \Pr(\text{missing}) = \alpha + \beta_1 diet + \beta_2 race_1 + \beta_3 race_2 \\ + \beta_4 grade + \beta_5 ldistW2. \end{aligned} \tag{2}$$

For missing-data models 1 and 2, we fixed the coefficients of this logistic regression at $\alpha = 3$, $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1$, $\beta_4 = 0.2$, and $\beta_5 = 0.3$, chosen to create a substantial level of association between variables and whether an observation was missing, and a reasonable amount of missingness. For example, $\beta_1 = 1$ corresponds to an odds ratio of 2.7, suggesting that girls who had dieted in the previous 7 days had more than double the odds of not responding as girls who had not dieted. The missing-data mechanism led to approximately 33% of values being missing for each variable, with both models 1 and 2 having data missing at random, since the missing-data mechanism depends only on fully observed variables. In model 3, we set $\beta_1 = 0$ to remove dependence of the missingness on diet (itself subject to missingness in this model), in order to focus on imputation of diet without the complication of data being missing not at random.

Analysis methods

For each simulated data set, with missing data imposed according to the mechanisms described, we estimated the regression model of interest (equation 1) using complete-case analysis—that is, restricting the data to cases where all required variables were observed, and using multiple imputation, performed with MVNI or FCS.

MVNI was performed using a Stata implementation of Schafer's NORM program (26). Imputed health and fitness scores were rounded to the nearest value (range, 1–5). Similarly, the binary diet variable was imputed on a continuous scale and rounded to 0 or 1 by either simple or "adaptive" rounding, under which the cutoff to distinguish between 0 and 1 is based on a normal approximation to the binomial distribution, making use of the marginal proportion of 0's and 1's in the observed data (27). FCS was carried out using the *ice* command in Stata (13), with the default number of 10 cycles. The diet variable was imputed using logistic regression, and health and fitness were imputed using ordinal (proportional odds) logistic regression, with grade, health, and fitness included as categorical predictors in each of the regression models.

Both approaches to imputation assume normality for continuous variables, and we considered a range of methods to mitigate the potential effects of nonnormality in the highly skewed distress measure, which was sometimes missing at wave I. The default option was simply to ignore the nonnormality and use the raw values in the imputation model. Imputed values less than or equal to 0 were replaced with the smallest observed value in the sample, and values greater than 3 were truncated at 3. A second approach for both MVNI and FCS was to use a log transformation, as in the analysis model, again truncating at 3 on the raw scale. A further alternative was to use a log transformation with an offset such that the observed values of the transformed variable had zero skewness; that is, we imputed $u = \ln(\pm x - k)$, choosing k and the sign of x so that u had zero skewness. This is convenient to implement in Stata using the *lnskew0* command, and we refer to this as the "log-skew0" transformation. Again, imputed values were truncated at the smallest observed value in the sample and at 3 on the raw scale. With each of these approaches, we first transformed the observed values of distress at both wave I and wave II in the sample, performed imputation for unobserved values (at wave I), and then back-transformed to obtain imputed values on the original scale before proceeding to the regression analysis of interest. The target analysis used the simple log transformation for both distress variables, as would be likely in practice, to maximize interpretability of the regression coefficients.

Within the FCS framework, we also compared results obtained under the method of prediction matching, as described above.

For all imputation methods, all covariates used in the analysis model (equation 1), as well as the outcome, were included in the imputation model to ensure the maximum recovery of information about associations of interest (28). Twenty imputed data sets were used for each analysis, with inferences for the regression coefficients being obtained by combining the results over the imputed data sets using Rubin's rules (3).

Comparisons

We compared the results of the complete-case analysis with results obtained using MVNI and FCS, with each of these methods having 3 variants according to the specification used for baseline distress (raw scale, log transformation, and log-skew0 transformation), and we also performed a further comparison with FCS using prediction matching. We assessed the properties of the regression coefficient estimates by analyzing results from 1,000 simulated data sets. We report the bias (average difference between estimate and population value) and the coverage of the estimated 95% confidence interval for each coefficient estimate. Based on the simulation sample size of 1,000, the standard error of the estimated coverage was 0.69% for a true coverage of 95%, implying that the estimated coverage should lie within the range 93.6%–96.4% (with 95% probability). We also report the average (estimated) standard error of each coefficient estimate, to give an indication of gains in precision due to recovery of information via imputation. All analyses were conducted in Stata, release 10 (29).

RESULTS

Table 1 gives a summary of the variables in the complete synthetic data set. Girls who dieted tended to be more distressed at wave I, to be in a slightly higher grade, and to have higher self-reported health and fitness but had similar levels of distress at wave II.

Under missing-data model 1 (Table 2), clear bias in the main effect of interest, for dieting, was apparent under complete-case analysis, as expected, and all imputation methods substantially ameliorated this bias. The method used to handle the missing distress variable had no substantial effect on the diet estimate, despite slightly larger bias when no transformation was used. All imputation approaches led to increased precision in the diet effect, due to the recovery of information from cases for whom the baseline distress information was missing.

The effect of the continuous predictor, emotional distress at wave I, was well estimated in the complete-case analysis, but when multiple imputation was used without due attention to the skewness problem, substantial biases were observed, leading to poor coverage. With both FCS and MVNI, use of log transformation only partially resolved the bias, but use of the log-skew0 scale produced much better results, as did prediction matching under FCS. Figure 1 displays imputed values obtained under each approach, showing that imputations using the raw data and the simple log transformation produced anomalies in the tails of the distribution for this variable.

The estimates for the categorical predictors of health and physical fitness were generally accurate for all analyses, even complete-case analysis, with similar results being obtained using FCS or MVNI irrespective of the transformation used for distress. Biases were minimal because these covariates were not associated with the missingness mechanism, but precision was gained by using multiple imputation.

Introducing missingness in the health and physical fitness variables (model 2) did not change the general pattern of the results (Table 3), although the bias in the diet effect was more pronounced with complete-case analysis, and relative gains in precision under multiple imputation were greater because of the higher proportion of cases with missing data. Although bias was reduced and coverage was slightly improved for the diet effect under multiple imputation as compared with complete-case analysis, there was slight undercoverage with both FCS and MVNI. Addressing the skewness in the distress covariate again clearly improved the associated coefficient estimate and in this case also led to slight reductions in bias for the health and fitness estimates, although coverage was not as consistent for these variables as under the more conservative complete-case approach.

Table 4 shows the results obtained when missingness was also introduced in the diet variable, with dependence of the missingness mechanism on diet removed (missing-data model 3). We again saw small biases in the diet effect with MVNI and FCS, irrespective of the transformation used for distress. As with missing-data models 1 and 2, there was large bias and poor coverage in the distress estimate when imputations were carried out on the raw or log scale for both FCS and MVNI, with the log-skew0 transformation and prediction matching only partially ameliorating the problem. Use of adaptive rounding under MVNI for the binary diet variable appeared to slightly improve coverage for all of the covariate effects.

The same general pattern of results was seen when simulations were carried out in the original data set, in which the diet effect was effectively null (results not shown).

DISCUSSION

Our primary aim was to investigate whether the MVNI approach to multiple imputation, with its reliance on an unrealistic multivariate normal modeling assumption, was inferior to the more flexible FCS approach in an analysis typical of those carried out in epidemiologic research. We found no evidence that MVNI performed less well; and in fact, somewhat surprisingly, it produced slightly better coverage than FCS, especially when used with adaptive

Table 1. Synthetic Data Set[a] Used for a Simulation Study Comparing 2 Methods of Multiple Imputation

	No Dieting	Dieting	Total
No. of girls	778,912	192,415	971,327
Covariate data			
Median emotional distress[b] at wave I (IQR)	0.58 (0.32–0.89)	0.63 (0.37–0.95)	0.58 (0.32–0.89)
Race, %			
White	60	65	61
Black	25	18	23
Nonblack Hispanic	15	17	16
Grade, %			
7	17	13	16
8	17	14	16
9	21	21	21
10	23	26	23
11	22	26	23
Mean grade (SD)	9.17 (1.39)	9.39 (1.34)	9.21 (1.38)
Overall health, %			
1	25	19	24
2	40	40	40
3	28	31	28
4	7	10	8
5	0	1	0
Mean health (SD)	2.18 (0.91)	2.34 (0.91)	2.22 (0.921)
Physical fitness, %			
1	21	11	19
2	47	43	46
3	22	30	23
4	10	15	11
5	1	2	1
Mean fitness (SD)	2.25 (0.93)	2.55 (0.93)	2.31 (0.94)
Outcome data			
Median emotional distress[b] at wave II[c] (IQR)	0.55 (0.29–0.90)	0.54 (0.29–0.88)	0.55 (0.29–0.90)

Abbreviations: IQR, interquartile range; SD, standard deviation.

[a] A synthetic population was created to resemble data from the US National Longitudinal Study of Adolescent Health (25). The primary covariate was a binary indicator for whether or not a girl had dieted in the previous 7 days at wave I.

[b] Possible range, 0–3.

[c] Outcome after shifting of the diet effect.

rounding for the binary diet variable. The slight undercoverage seen for several parameters under both approaches to multiple imputation in models 2 and 3 is difficult to explain, particularly given the small biases. It appears to be due mainly to variation in the standard errors across simulations and to negative correlations between the estimates and their standard errors. It is unclear whether this pattern is likely to arise in general or whether it is specific to these simulation models.

The main advantage of MVNI in practice is the ease of specification of the imputation model, but many would-be users are concerned about the unrealistic nature of the multivariate normal assumption. On the other hand, FCS often requires more effort in model specification, since a separate regression model must be fitted for each variable in the imputation model (14). Although the conditional regressions can be automatically specified in problems with a reasonably small number of variables, this becomes more difficult with large data sets, especially when there are many variables subject to missingness. In addition, there is the theoretical problem of potential incompatibility between the conditional specifications for each variable that is

Table 2. Parameter Estimates Obtained Under Missing-Data Model 1: Missing Values in Baseline Emotional Distress[a]

Imputation Method	Distress Transformation	Diet (β_1 = −0.101)			Emotional Distress (β_2 = 0.554)			Health (β_6 = 0.042)			Physical Fitness (β_7 = 0.056)		
		Bias	SE[b]	Coverage	Bias	SE	Coverage	Bias	SE	Coverage	Bias	SE	Coverage
CCA	None	−0.039	0.090	0.923	−0.008	0.043	0.941	0.001	0.039	0.950	0.003	0.038	0.945
FCS	None	0.012	0.069	0.955	−0.086	0.044	0.484	0.011	0.034	0.954	0.010	0.032	0.955
FCS	Log	0.007	0.069	0.958	−0.044	0.039	0.769	0.007	0.033	0.953	0.003	0.032	0.948
FCS	Log-skew0	0.004	0.068	0.948	−0.012	0.042	0.942	0.003	0.033	0.947	0.001	0.031	0.938
FCS-M	None	0.003	0.068	0.955	−0.013	0.039	0.916	0.002	0.033	0.963	0.002	0.031	0.950
MVNI	None	0.010	0.069	0.959	−0.079	0.043	0.545	0.009	0.032	0.950	0.009	0.032	0.956
MVNI	Log	0.006	0.069	0.951	−0.038	0.039	0.803	0.006	0.032	0.950	0.002	0.032	0.944
MVNI	Log-skew0	0.003	0.068	0.946	−0.005	0.040	0.943	0.002	0.032	0.952	<0.001	0.031	0.939

Abbreviations: CCA, complete-case analysis; FCS, fully conditional specification; FCS-M, fully conditional specification fitted using prediction matching; MVNI, multivariate normal imputation; SE, standard error.

[a] Average of 667 observations with complete data.

[b] Average (estimated) standard error across the 1,000 data sets.

imputed, although it is unclear how often this might lead to problems in practice (14). Previous comparisons between the 2 approaches have produced mixed results. Yu et al. (16) and van Buuren (14) both showed that imputations under the multivariate normal assumption may not reflect the distribution of observed values and concluded that a conditional approach was more reliable. However, we have shown that a binary covariate can be imputed adequately using MVNI and that failure to reproduce the full distributional shape may not adversely affect inferences for regression coefficients of interest. Similarly, Demirtas et al. (30) found that MVNI performed well even if multivariate normality did not hold.

Clearly, the ease of specification associated with MVNI may bring risks if results are highly sensitive to lack of normality, although we did not observe this in relation to the binary covariate of primary interest in our analysis. However, we did observe substantial sensitivity to nonnormality in a highly predictive continuous covariate—a problem which affected the FCS approach as well. We explored the use of the log-skew0 transformation to produce a symmetric distribution in the imputation model and showed in our example that it dramatically improved the accuracy of estimation under both FCS and MVNI. An alternative option with FCS is prediction matching, which can deal with more general nonnormality than right-skewness. Although it was effective in this example, prediction matching may not always be reliable, especially if the study sample is small or there are strong dependencies between missingness and covariates (16, 31). Our analysis illustrated the important point that a different transformation (in this case, log-skew0) can be used in the imputation model than in the analysis model and that this can improve the results. A further extension of the log-skew0 approach that would be worth exploring is the use of nonparametric normalizing transformations—for example, applying the inverse normal distribution function to the observed order statistics. A similar approach has been applied to normalize binary data (32).

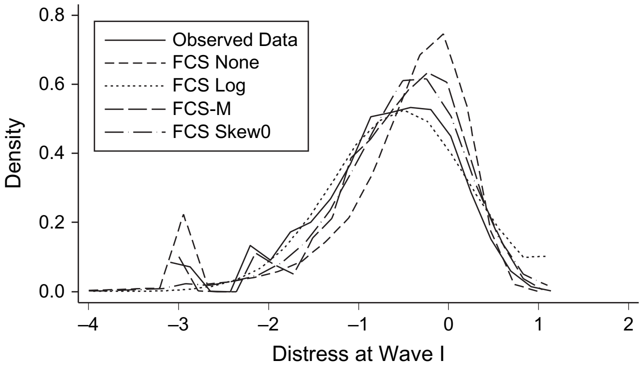

Figure 1. Distribution (kernel density) of observed data and values imputed under fully conditional specification (FCS) for various transformations of emotional distress at wave I, from missing-data model 1 (see Methods section in text). Results were based on 20 simulated data sets and 20 sets of imputations for each imputation method. Emotional distress is shown on the natural log scale. FCS-M, fully conditional specification fitted using prediction matching.

An advantage of the FCS approach is the natural handling of both ordinal and nominal variables. In this example, ordinal variables were imputed on the continuous scale and rounded to the required categories postestimation under MVNI, although there is ongoing research on how to categorize or whether it is necessary. Our analysis confirmed the benefits of adaptive rounding for imputing a binary variable, as proposed by Bernaards et al. (27). An extension of this method to calibrate imputed categorical variables may be valuable (33). We did not consider strictly categorical (nominal) variables in our analysis, although a suggested approach under MVNI is to use a set of dummy variables (34). Further investigation of issues associated with the imputation of categorical variables is needed.

Beyond the comparison between MVNI and FCS, our results provided another clear illustration of the potential gains from multiple imputation over complete-case analysis, when the missing data mechanism is missing at random.

Table 3. Parameter Estimates Obtained Under Missing-Data Model 2: Missing Values in Baseline Emotional Distress, Health, and Physical Fitness[a]

Imputation Method	Distress Transformation	Diet ($\beta_1 = -0.101$)			Emotional Distress ($\beta_2 = 0.554$)			Health ($\beta_6 = 0.042$)			Physical Fitness ($\beta_7 = 0.056$)		
		Bias	SE[b]	Coverage	Bias	SE	Coverage	Bias	SE	Coverage	Bias	SE	Coverage
CCA	None	−0.093	0.126	0.870	−0.016	0.052	0.942	0.002	0.049	0.947	0.007	0.048	0.938
FCS	None	0.001	0.070	0.931	−0.088	0.043	0.453	0.019	0.036	0.909	0.015	0.035	0.918
FCS	Log	<0.001	0.070	0.919	−0.044	0.040	0.760	0.009	0.036	0.934	0.005	0.035	0.932
FCS	Log-skew0	−0.004	0.069	0.931	−0.017	0.041	0.900	0.013	0.036	0.925	0.006	0.035	0.938
FCS-M	None	−0.007	0.069	0.923	−0.016	0.039	0.897	0.011	0.036	0.929	0.007	0.035	0.931
MVNI	None	0.002	0.070	0.959	−0.080	0.043	0.554	0.011	0.038	0.953	0.010	0.037	0.945
MVNI	Log	0.002	0.070	0.919	−0.037	0.040	0.816	0.001	0.038	0.919	<0.001	0.037	0.927
MVNI	Log-skew0	−0.003	0.069	0.927	−0.008	0.041	0.912	0.005	0.037	0.929	0.001	0.036	0.920

Abbreviations: CCA, complete-case analysis; FCS, fully conditional specification; FCS-M, fully conditional specification fitted using prediction matching; MVNI, multivariate normal imputation; SE, standard error.

[a] Average of 470 observations with complete data.

[b] Average (estimated) standard error across the 1,000 data sets.

In an additional analysis (not shown), we retained the dependence on diet in the third missing-data model, thus introducing an element of missingness not at random, and found substantially similar results. The imputation-based analysis readily corrected bias in the diet estimate and substantially improved precision in most of the covariate effects. Importantly, these gains mainly related to the recovery of cases for whom covariate values of interest were observed, not to the direct imputation of missing values for the target covariate of interest. However, such gains are not guaranteed, and they depend on a proper understanding of the missing-data mechanism as well as appropriate implementation of multiple imputation. For example, we showed that failure to impute on an appropriate scale could introduce bias that was not present in a complete-case analysis.

It is always difficult to draw general conclusions from a single simulation study, but we believe this study provided a good setting for a comparison between MVNI and FCS. The simulations were designed to be both realistic, in terms of data structure and a moderate degree of complexity, and demanding, by including substantial nonnormality, a high frequency of missing data, and a strong dependency of the missing-data process on variables in the analysis model. Undoubtedly, further exploration in real data sets, where of course the missing-data model is unknown, and in other simulation models would be useful. Further extensions

Table 4. Parameter Estimates Obtained Under Missing-Data Model 3: Missing Values in Baseline Emotional Distress, Health, Physical Fitness, and Diet[a]

Imputation Method	Distress Transformation	Diet ($\beta_1 = -0.101$)			Emotional Distress ($\beta_2 = 0.554$)			Health ($\beta_6 = 0.042$)			Physical Fitness ($\beta_7 = 0.056$)		
		Bias	SE[b]	Coverage	Bias	SE	Coverage	Bias	SE	Coverage	Bias	SE	Coverage
CCA	None	−0.006	0.113	0.951	−0.020	0.058	0.936	0.001	0.055	0.945	0.006	0.054	0.938
FCS	None	0.008	0.076	0.932	−0.077	0.042	0.552	0.010	0.037	0.931	0.009	0.036	0.924
FCS	Log	0.013	0.076	0.919	−0.035	0.039	0.818	−0.001	0.037	0.935	−0.001	0.036	0.920
FCS	Log-skew0	0.003	0.075	0.930	−0.010	0.040	0.920	0.001	0.036	0.935	<0.001	0.035	0.920
FCS-M	None	0.003	0.075	0.934	−0.012	0.038	0.902	0.003	0.036	0.922	0.002	0.035	0.928
MVNI	None	0.021	0.074	0.957	−0.073	0.043	0.601	0.010	0.036	0.942	0.008	0.036	0.945
MVNI	Log	0.024	0.074	0.926	−0.032	0.039	0.833	−0.001	0.036	0.934	−0.002	0.036	0.926
MVNI	Log-skew0	0.016	0.073	0.939	−0.005	0.040	0.924	0.001	0.036	0.932	−0.001	0.035	0.923
MVNI[c]	Log-skew0	−0.019	0.088	0.959	−0.007	0.041	0.948	0.003	0.037	0.953	−0.001	0.036	0.953

Abbreviations: CCA, complete-case analysis; FCS, fully conditional specification; FCS-M, fully conditional specification fitted using prediction matching; MVNI, multivariate normal imputation; SE, standard error.

[a] Average of 392 observations with complete data.

[b] Average (estimated) standard error across the 1,000 data sets.

[c] Adaptive rounding used for binary diet variable.

could be to explore the behavior of multiple imputation in the presence of a binary variable with very low prevalence and with a categorical variable for which a linear regression effect is not appropriate.

In summary, multiple imputation using either FCS or MVNI will often provide a useful and more reliable approach than complete-case analysis in the presence of missing data, and MVNI appears to perform well even in the presence of binary and ordinal variables. However, transformation of skewed variables to a symmetric distribution is strongly recommended to ensure reliable estimation of associations with that variable and to avoid the introduction of biases in inferences for other variables.

ACKNOWLEDGMENTS

Author affiliations: Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, The Royal Children's Hospital, Melbourne, Victoria, Australia (Katherine J. Lee, John B. Carlin); and Department of Paediatrics, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Victoria, Australia (Katherine J. Lee, John B. Carlin).

This work was partially supported by the Australian National Health and Medical Research Council (project grant 334336).

The authors are grateful to Dr. John Galati (Murdoch Childrens Research Institute, Melbourne, Victoria, Australia) for the development of Stata programs used in this work and to Dr. Joe Schafer (Pennsylvania State University, State College, Pennsylvania) for providing his synthetic population data set, on which these simulations were based.

Conflict of interest: none declared.

REFERENCES

1. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
2. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. *Am J Epidemiol*. 2008;168(4): 355–357.
3. Rubin D. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, Inc; 1987.
4. Schafer JL. *Analysis of Incomplete Multivariate Data*. London, United Kingdom: Chapman & Hall Ltd; 1997.
5. Carlin JB, Galati JC, Royston P. A new framework for managing and analysing multiply imputed data sets in Stata. *Stata J*. 2008;8(1):49–67.
6. Royston P, Carlin JB, White IR. Multiple imputation of missing values: new features for mim. *Stata J*. 2009;9(2): 252–264.
7. SAS Institute Inc. The MI procedure. In: *SAS/STAT 9.1 User's Guide*. Cary, NC: SAS Institute Inc; 2004:2509–2606.
8. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *Am Stat*. 2007;61(1):79–90.
9. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*. 1999;18(6):681–694.
10. Raghunathan TE, Siscovick DS. A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Appl Stat*. 1996;45(3):335–352.
11. Raghunathan TE, Lepkowski JM, Van Hoewyk J, et al. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol*. 2001; 27(1):85–95.
12. Royston P. Multiple imputation of missing values. *Stata J*. 2004;4(3):227–241.
13. Royston P. Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. *Stata J*. 2007;7(4):445–464.
14. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*. 2007;16(3):219–242.
15. van Buuren S, Brands JPL, Groothuis-Oudshoorn CGM, et al. Fully conditional specification in multivariate imputation. *J Stat Comput Simul*. 2006;76(12):1049–1064.
16. Yu LM, Burton A, Rivero-Arias O. Evaluation of software for multiple imputation of semi-continuous data. *Stat Methods Med Res*. 2007;16(3):243–258.
17. Schafer JL. *Software for Multiple Imputation* [software]. State College, PA: Department of Statistics, Pennsylvania State University; 1999. (http://www.stat.psu.edu/~jls/misoftwa.html). (Accessed April 6, 2009).
18. Stata Corporation. *Stata Statistical Software: Release 11* [software]. College Station, TX: Stata Corporation; 2009.
19. Choi KH, Hoff C, Gregorich SE, et al. The efficacy of female condom skills training in HIV risk reduction among women: a randomized controlled trial. *Am J Public Health*. 2008;98(10): 1841–1848.
20. Seitzman RL, Mahajan VB, Mangione C, et al. Estrogen receptor alpha and matrix metalloproteinase 2 polymorphisms and age-related maculopathy in older women. *Am J Epidemiol*. 2008;167(10):1217–1225.
21. Gelman A, Hill J, Yajima M, et al. *mi: Missing Data Imputation and Model Checking* [software]. Vienna, Austria: R Project for Statistical Computing; 2009. (http://cran.r-project.org/web/packages/mi/index.html). (Accessed June 14, 2009).
22. Joseph JG, El-Mohandes AA, Kiely M, et al. Reducing psychosocial and behavioral pregnancy risk factors: results of a randomized clinical trial among high-risk pregnant African American women. *Am J Public Health*. 2009;99(6):1053–1061.
23. Stuart E, Azur M, Frangakis C, et al. Multiple imputation with large data sets: a case study of the children's mental health initiative. *Am J Epidemiol*. 2009;169(9):1133–1139.
24. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. 2nd ed. New York, NY: John Wiley & Sons, Inc; 2002.
25. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychol Methods*. 2008;13(4):279–313.
26. Galati JC, Carlin JB. *INORM: Stata Module to Perform Multiple Imputation Using Schafer's Method* [software]. Chestnut Hill, MA: Department of Economics, Boston College; 2008. (http://ideas.repec.org/c/boc/bocode/s456966.html). (Accessed April 8, 2009).
27. Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med*. 2007;26(6):1368–1382.

28. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behav Res.* 1998;33(4):545–571.
29. Stata Corporation. *Stata Statistical Software: Release 10* [software]. College Station, TX: Stata Corporation; 2007.
30. Demirtas H, Freels SA, Yucel RM. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *J Stat Comput Simul*. 2008;78(1):69–84.
31. Royston P. Multiple imputation of missing values: update of ice. *Stata J*. 2005;5(4):527–536.
32. Nevalainen J, Kenward MG, Virtanen SM. Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. *Stat Med*. 2009; 28(29):3657–3669.
33. Yucel RM, He Y, Zaslavsky AM. Using calibration to improve rounding in imputation. *Am Stat*. 2008;62(2):125–129.
34. Allison PD. *Missing Data.* (Quantitative Applications in the Social Sciences). Thousand Oaks, CA: Sage Publications, Inc; 2002.