Published in final edited form as:

Stat Med. 2011 December 20; 30(29): 3447–3460. doi:10.1002/sim.4355.

Gaussian-based routines to impute categorical variables in health surveys

Recai M. Yucel^{a,*†}, Yulei He^b, and Alan M. Zaslavsky^b

^aDepartment of Epidemiology and Biostatistics, School of Public Health, University at Albany, SUNY, One University Place, Rensselaer, NY 12144-3456, USA

^bDepartment of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA, USA

Abstract

The multivariate normal (MVN) distribution is arguably the most popular parametric model used in imputation and is available in most software packages (e.g., SAS PROC MI, R package norm). When it is applied to categorical variables as an approximation, practitioners often either apply simple rounding techniques for ordinal variables or create a distinct ‘missing’ category and/or disregard the nominal variable from the imputation phase. All of these practices can potentially lead to biased and/or uninterpretable inferences. In this work, we develop a new rounding methodology calibrated to preserve observed distributions to multiply impute missing categorical covariates. The major attractiveness of this method is its flexibility to use any ‘working’ imputation software, particularly those based on MVN, allowing practitioners to obtain usable imputations with small biases. A simulation study demonstrates the clear advantage of the proposed method in rounding ordinal variables and, in some scenarios, its plausibility in imputing nominal variables. We illustrate our methods on a widely used National Survey of Children with Special Health Care Needs where incomplete values on race posed a valid threat on inferences pertaining to disparities.

Keywords

missing data; multiple imputation; nominal data imputation; software; rounding

1. Motivating example: children with special health care needs

The National Survey of Children with Special Health Care Needs (NSCSHCN) provides a unique opportunity to investigate issues related to children with special health care needs (CSHCN) and their families. Among many substantively important uses of this survey, our particular problem pertains to drawing inferences on the correlates of unmet health care and family service needs for all CSHCN. Following a broad definition released by the US Maternal and Child Health Bureau [1], CSHCN are those children who require services beyond that required by children generally because of a chronic physical, developmental, behavioral, or emotional condition. Using this definition, approximately 12.8% of children in the USA had a special health care need in 2001 [2]. Previous studies [3, 4] show that CSHCN use more health care services and that their families experience a variety of

consequences of caring for a child with SHCN such as lost employment and increased financial burden.

Our proposed method is largely motivated by the need to make full use of observed data in the NSCSHCN. If one were to only proceed with complete-case only analysis, only 68% of the sample would be analyzed. As documented by many researchers, such analyses have potential undesired inferential properties including bias and distorted estimates on the uncertainty measures. An increasingly popular inferential method to accomplish this is multiple imputation (MI) [5]. Briefly, MI is a simulation-based inferential tool operating on $M > 1$ ‘completed’ data sets, where the missing values are replaced by random draws from their respective predictive distributions (e.g., posterior predictive distribution of missing data). These M versions of completed data are then analyzed by standard complete-data methods, and the results are combined into a single inferential statement using rules to yield estimates, standard errors, and p -values that formally incorporate the missing-data uncertainty into the modeling process [5]. The key ideas and the advantages of MI are given by [5, 6].

From a practitioner’s perspective, MI is a flexible and convenient solution. Growing availability of MI software (e.g., SAS PROC MI) [9] has contributed greatly to its popularity. The underlying normality assumption, however, is probably the single most violated assumption of this procedure. In most health surveys, for example, most items subject to incomplete data are the items measured categorically using either nominal or ordinal scales. As demonstrated in the forthcoming sections, utilizing incorrect distributional assumptions can pose a great inferential threat. Our methods aim to remove this threat with minimal adjustments to the current methods.

Our imputation methods are specifically developed for incompletely observed categorical variables. These methods extend the calibration techniques developed by [7] for binary imputation to the ordinal and nominal variable imputation. Our methods are largely motivated by practical purposes and allow practitioners to adopt them with minimal programming. The unique contribution of these techniques is the ability to employ Gaussian-based (or any working distribution) imputation techniques which avoid the common computational problem due to dimensionality and/or scarcity ([6, 7]).

The substantive focal point of our methodology is the study of racial disparities within the context of unmet needs for all CSHCN and for CSHCN with severe conditions. Understanding the disparities in these outcomes is an important topic in health services research, and incomplete values on the key race variable introduce analytical difficulties and create a potential source of statistical bias if no sound action is taken. Descriptive summaries of the selected survey items from the NSCSHCN can be seen in Table I.

One of the most frequently asked questions in the analysis of incomplete data pertains to implications of conducting analyses with only complete cases. To have empirical insight about such implications, practitioners can analyze the missingness patterns. In our application, for example, disproportionate distributions of key variables corresponding to observed and missing cases (e.g., CSHCN with severe conditions with race) raise a valid concern for misleading conclusions under an analysis ignoring incomplete cases. Biased inferences as well as under-powered conclusions are the usual adverse outcomes of such analyses.

Although missing data exist in most of the NSCSHCN variables, to illustrate our calibration technique, we focus on a single variable, the race variable, which is measured on a nominal scale. Although relatively low levels of missing data are found on race variable (around 4.5%), those who are missing on race had significantly high rates of unmet health needs and

CSHCN with severe conditions. Table I summarizes descriptive statistics and compares cases with missing race values, which do not resemble the same distribution as the cases with complete race values. Among the most notable differences, we see that cases with missing values have lower levels of maternal education, higher poverty levels, and higher rates of public/private and only private insurance. A significantly higher rate of Hispanics is also seen in the cases with missing race information. These issues pose real analytical problems for health services researchers who face incomplete race data in disparity studies. In addition, as Section 5 elaborates more, our methods can easily accommodate arbitrary missing values in other types of variables.

The remainder of this article is organized as follows. In the next section, we summarize the software and previous methods on rounding under multivariate normal (MVN) distribution. Section 3 then describes our rounding strategy on the basis of the calibration of the marginal distribution and states how it can be implemented for nominal and ordinal variables. Section 4 summarizes the results from a simulation study assessing the performance of our method. Section 5 demonstrates an application using the NSCSHCN to explore the correlates of self-reported CSHCN with severe conditions. Finally, Section 6 discusses the strengths and weaknesses of this approach.

2. Need for rounding in Gaussian-based multiple imputation

With the growing availability of software, MI under the MVN distribution has emerged as a popular inferential tool in the analysis of incomplete data. Among most commonly used software packages are the missing library of S-Plus [8], the norm library of R [6], and SAS PROC MI [9]. These software packages commonly implement specialized computational techniques for sampling from the implied posterior predictive distribution under MVN and a set of priors. Sampled data points are regarded as multiple imputations and allow the subsequent analyses to continue with these ‘completed’ datasets. After the analyses are conducted across these datasets, simple rules defined by [5] are applied to combine the estimates and quantify the impact of missing data on the final uncertainty measures.

The plausibility of the normality assumption is a major concern in adapting the software for MI under MVN. In many real-life problems with categorical variables measured on a nominal scale, the MVN assumption is clearly violated. In previous studies, two distinct solutions have been proposed to remedy this problem. The first solution develops a strict categorical model used as a basis for imputation (see [6], Chap. 6). Because of its practical as well as computational limitations on the number of variables, this solution has been of limited use to practitioners. The second solution has mostly been limited to cases where the binary variable has missing values and modified existing techniques for continuous data to impute the binary variables [7, 10]. Because we focus on using the existing MVN-based procedures to impute a larger class of nominal and ordinal variables, the following discussion is limited to the latter approach.

Modification of the existing techniques under MVN to impute categorical data varies greatly and is fairly easy to implement. In most practices of rounding, imputed values for binary or ordinal variables are rounded off to the nearest observed values ([6], Chap. 5). This approach can lead to biased estimates (e.g., means or correlations) when the binary or ordinal variables have distributions that are far from symmetry or oddly shaped (e.g., multimodal). Note that this distribution is used as a basis for an approximation (albeit a non-sensical one for some ordinal and all nominal variables) to the posterior predictive distribution of missing categorical variables. The disadvantages of such methods are likely to worsen as the amount of missing data increases. Horton *et al.* [11] evaluated the imputation of a Bernoulli random variable by rounding a continuous imputation to the

nearest value of 0 or 1 (i.e., by a cutoff value of 1/2). They found that this caused substantial bias and suggested a normal imputation of the binary variable without rounding to obtain an unbiased estimate of the mean (probability of success). Although unbiasedness is an important property, the latter study's strategy of no rounding does not impute data of the original type as might be desired by practitioners whose intended complete-data analysis is tailored to binomial data. Bernaards *et al.* [10] evaluated the robustness of the MVN approximation for the imputation of incomplete binary data and suggested a rule for calculating a cutoff value on the basis of a normal approximation to the binomial distribution. More recently, [12] reviewed and provided suggestions on rounding strategies tailored towards the goal of the analysts.

As in the binary case, naive rounding for ordinal values can potentially lead to substantial biases, and other methods suffer from the similar shortcomings stated in the preceding paragraphs. Further, the current rounding practices under MVN cannot be utilized for imputing nominal variables. One approach is to create a qualitatively different 'missing' category, limiting the scope of the inferences and often not recommended [13].

The main motivation of our work was to offer a simple but principled analytical solution for the analysis of incompletely observed variables. These solutions are centered around MVN-based MI software and develop rules for rounding nominal variables, extending our recent work ([7], from hereon referred to as YHZ) where imputed values are calibrated to resemble the observed distribution. Another important contribution of this paper is the demonstration of the calibration on the NSCSHCN. It should be noted that, although the MVN is the underlying assumption of the commonly used software, the main assumption here is the existence of the working method to impute continuous values for the nominal variables; hence, the methods described here do not assume a particular distribution.

3. Calibration-based rounding strategy

3.1. Notations and assumptions

Let Y denote the nominal variable of interest subject to missingness, and let $1, 2, \dots, G$ denote the values assumed by Y . Further, let X denote the covariates (univariate or multivariate) in the imputation model, where Y consists of observed and missing values, Y_{obs} and Y_{mis} , respectively. We assume that X either is complete or has been completed by a plausible imputation method. The goal in inference by MI is to replace Y_{mis} by random draws from its posterior predictive distribution $P(Y_{\text{mis}} | Y_{\text{obs}}, X)$.

Under MVN-based imputation procedure, for example, X and Y are assumed to follow a MVN distribution. Typically, a data augmentation scheme is implemented by widely used software (e.g., SAS PROC MI, with the 'mcmc' option). This computational algorithm proceeds as a Gibbs sampler: First, a value of missing data from the conditional predictive distribution of Y_{mis} is drawn from the posterior predictive distribution, $P(Y_{\text{mis}} | Y_{\text{obs}}, X, \theta^{(t)})$, where $\theta^{(t)}$ denotes the current value for the mean vector and variance-covariance matrix.

Then, conditioning on $Y_{\text{mis}}^{(t+1)}$, a new value of θ from its complete-data posterior,

$\theta^{(t+1)} \sim P(\theta | Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$, is drawn. Repeating these two steps under a starting value $\theta^{(0)}$

yields a stochastic sequence $\{\theta^{(t)}, Y_{\text{mis}}^{(t)} : t=1, 2, \dots\}$ whose stationary distribution is $P(\theta, Y_{\text{mis}} | Y_{\text{obs}}, X)$, and further the subsequences $\theta^{(t)}$ and $Y_{\text{mis}}^{(t)}$ have $P(\theta | Y_{\text{obs}}, X)$ and $P(Y_{\text{mis}} | Y_{\text{obs}}, X)$ as their respective distributions. For a reasonably large number of iterations, convergence to these stationary distributions is achieved. Because the complete-data likelihood is assumed to follow a MVN distribution, drawing from conditional distributions above is relatively straightforward. A simple regression of missing data on observed data in a given

missingness pattern for a given state of the unknown parameters can be performed to draw missing data, making the normality assumption a computationally attractive structure.

Another set of notation that is necessary to introduce pertains to the missingness mechanism. Following standard notation, let R be the indicator variable for an observation Y , with $R = 1$ for an observed Y and $R = 0$ for a missing Y . Under a missing at random (MAR) missingness mechanism (as assumed by the MVN-based methods summarized above), the probability that any data value is missing may depend on quantities that are observed but not on quantities that are missing: $P(R | Y_{\text{obs}}, Y_{\text{mis}}, X) = P(R | Y_{\text{obs}}, X)$. Under a missing completely at random (MCAR) mechanism, missingness is independent of both observed and missing values: $P(R | Y_{\text{obs}}, Y_{\text{mis}}, X) = P(R)$.

The notation below builds upon the working assumption that determines the form of this posterior predictive distribution from which the missing data are simulated. We let Y_c^* denote a variable imputed under a continuous model $P(Y_{\text{mis}} | Y_{\text{obs}}, X)$ which will be rounded off to a variable that takes values in the desired scale (e.g., ordinal or unordered nominal), denoted by Y^* . The rounding will proceed using a set of rules determined by calibration. In the binary case considered by YHZ, for example, this rule is given by

$$Y^* = \begin{cases} 1 & Y_c^* > c, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where c is the *cutoff* value determined to meet a criterion (e.g., mean of the observed distribution) of calibration.

3.2. Calibration idea

The main goal of the calibration is to create imputed values with a similar distribution to that of the observed values. The use of a multivariate distribution to establish this goal can also achieve the secondary goal, preserving relationships with other potentially important variables to the subsequent analyses. To implement a method with such properties, [7] simply duplicated and appended a copy to the original data, and Y is intentionally set to missing in the second copy. Then imputations Y^* of the missing copy of Y as well as the imputations for the missing data in the original data are generated. This procedure is illustrated in Figure 1 using a hypothetical binary sample in which 70% are coded as one (i.e., success probability is estimated to be 70%). The Y column shows the original data and the duplicate binary data where all values are set to missing. Using an imputation procedure (e.g., PROC MI or R norm package), the entire set (original and duplicate) is imputed. The duplicate copy now includes the imputed values for the observed cases as well as the missing cases. The calibration approach then uses the subset of imputed values corresponding to the observed cases to determine the cutoff. For computational ease, one can order these imputed values and choose the cutoff value (y_c) to achieve the same success rate (i.e., $\hat{p} = 0.70$) and apply it for rounding in the imputed values corresponding to the missing cases. Yucel *et al.* [7] demonstrated this method on a binary case and derived the asymptotic biases in estimating means and coefficients, which are largely tolerable in problems with modest missingness and in problems with strong ‘tail’ behavior (i.e., distributions with extremely high or low success probabilities).

In the context of developing rounding rules, the motivation of duplication is as follows.

Under a given imputation model, the sampled or imputed values of observed data, Y_{obs}^* , conditional on the covariates, X , are seen as coming from the same distribution generating Y_{obs} . Then the calibration proceeds with respect to a statistic $S(X, Y_{\text{obs}})$. In other words, we

can develop rounding rules under the assumption of $S(X, Y_{\text{obs}})$ because it is a realization from the distribution of the statistic evaluated under the imputation model, $S(X, Y_{\text{obs}}^*)$. For example, in the hypothetical example used in Figure 1, $S(\cdot)$ is the sample mean. In our example used in Section 5, $S(\cdot)$ will be the relative frequencies of the race variable. A natural choice in the problem of imputation of a nominal variable can be to make a procedure to have a property $S(X, Y_{\text{obs}}^*) = S(X, Y_{\text{obs}})$, where $S(\cdot)$ is the probability of categories or some aspect of its joint distribution with X . Our procedures consider $S(\cdot)$ to be the mean (i.e., probability).

3.3. Marginal calibration for ordinal variables

The strategy for extending the methodology of YHZ into ordinal variables uses the same principle: Obtain the set of cutoff values that will be used in rounding the imputed and calibrated $Y_{\text{mis},C}^*$ in such a way that, after rounding, Y_{obs} will have a marginal distribution consistent with the observed data Y_{obs} . The algorithm that implements this principle is given by

1. $X_{\text{dup}} = \{X, X\}$ and $Y_{\text{dup}} = \{Y_{\text{obs}}, Y_{\text{mis}}, Y_{\text{obs(dup)}}, Y_{\text{mis(dup)}}\}$ are the duplicated datasets; $Y_{\text{obs(dup)}}$ means Y_{obs} with observed values changes to missing.
2. Impute the missing values in Y_{dup} under MVN (or any chosen continuous distribution) and denote the imputed continuous values as $Y_{\text{mis},C}^*, Y_{\text{obs(dup),C}}^*$, and $Y_{\text{mis(dup),C}}^*$.
3. The cutoff values $c_1, c_2, \dots, c_G$ are determined so that the relative frequencies of each category in $Y_{\text{obs(dup)}}^*$, equals the relative frequency in Y_{obs} .
4. Use $c_1, c_2, \dots, c_G$ to round $Y_{\text{mis},C}^*$ to Y_{mis}^* :

$$Y_{\text{mis}}^* = g \text{ for } c_{g-1} \leq Y_{\text{mis},C}^* < c_g, \quad (2)$$

where $c_0 = -\infty$ and $c_G = \infty$.

Note that $c_1, c_2, \dots, c_G$ can be any value between the appropriate order statistics of $Y_{\text{obs(dup),C}}^*$. To reflect uncertainty in the practice of MI, one can draw the cutoff values from a uniform distribution defined in the appropriate interval given in (2).

3.4. Marginal calibration for nominal variables

When the variable of interest Y is an unordered nominal variable (e.g., race), even approximately, current techniques assuming MVN cannot be utilized. The calibration routine introduced in the preceding paragraphs can be used with slight modifications to reflect the unordered nature and proceeds with a set of dummy variables and rounding rules that are based on cumulative probabilities. The same algorithm of duplication, imposition of missingness, and calibration applies to each of the dummy variables in a sequential manner so that the imputation proceeds in a fashion that is mutually exclusive and collectively exhaustive:

1. First create a set of dummy variables in Y_{obs} indicating the underlying category: $I_{Y_{\text{obs}(i)}=g}$:

$$I_{Y_{\text{obs}(i)}=g} = \begin{cases} 1 & \text{if } Y_{\text{obs}(i)}=g, \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

This will thus turn the input for the MVN-based imputation procedure into a slightly different format:

$$X_{\text{dup}}, I_{Y_{\text{obs}(i)}=1}, I_{Y_{\text{obs}(i)}=1}^{\text{copy}}, \dots, I_{Y_{\text{obs}(i)}=g-1}, I_{Y_{\text{obs}(i)}=g-1}^{\text{copy}},$$

where the superscript ‘copy’ indicates the duplicate copy of the indicators which will be missing in the second copy.

2. Impute the missing values in the original copy of Y (i.e., $I_{Y_{\text{obs}}}$) as well as in the $\left\{ I_{Y_{\text{obs}(i)}=g}^{\text{copy}} \right\}_{g=1}^G$ under MVN (or any chosen continuous distribution).
3. Compute the cutoff values $c_1, c_2, \dots, c_{G-1}$ in a sequential fashion (case where $G = 3$ is given in the following):
 - Choose c_1 so that when $Y_{c, I_{Y_{\text{obs}(i)}=1}}^*$ (imputed value under continuous model corresponding to the indicator for $Y = 1$) is rounded, the relative frequency of the first category in the duplicate copy matches the relative frequency of $Y_{\text{obs}} = 1$ (note that $Y_{c, I_{Y_{\text{obs}(i)}=1}}^* > c_1$ corresponds to the indicator for $Y = 1$).
 - For those cases with $Y_{c, I_{Y_{\text{obs}(i)}=1}}^* \leq c_1$, choose c_2 so that when $Y_{c, I_{Y_{\text{obs}(i)}=2}}^*$ (imputed value under continuous model corresponding to the indicator for $Y = 2$) is rounded, the relative frequency of the second category in the duplicate copy matches the relative frequency of $Y_{\text{obs}} = 2$.

More generally, we compute the cutoff value for $Y_{c, I_{Y_{\text{obs}(i)}=g}}^*$ in such a way that a fraction of those with $Y_i = g$ in the duplicate copy is equal to the fraction in the observed data. However, we cannot choose the cutoff values independently as the corresponding dummy variables may be imputed as 1, to avoid an identifiability problem. Finally, to reflect the uncertainty due to the imputation process (or sampling from the approximated posterior predictive distribution of missing data), we draw the cutoff values for a uniform distribution defined in the interval within the appropriate cutoff values. Note that this strategy is not affected by the sequence followed to set up the dummy variables as the goal is to match frequencies in the categories defined by the dummies.

4. Simulation study

We assessed the operational characteristics of our methods in a simulation study following a common practice of MI. It consists of data generation, imposing a missingness mechanism, performing MI inference under a hypothetical analyst’s model, and evaluating this inference under conventional (rounding to nearest observed integer) and calibration-based rounding.

4.1. Design

We first created an ordered or ordinal categorical variable dependent on a continuous variable. Specifically, we simulated two random variables X and U under

$$\begin{aligned} X &\sim N(0, 1) \\ U | X &\sim N(-1.6 + 0.2X, 4), \end{aligned}$$

and we grouped the data into categories using the following percentiles, (0.1, 0.2, 0.3, 0.55), creating the ordinal variable denoted by Y . In other words, the bottom 0th to 10th percentile values are category 1, 10th to 20th values are category 2, and so on. These probabilities were chosen specifically to obtain a highly skewed Y to emphasize the disadvantages of current practices. Missing values on Y were introduced under MAR. To do this, we simulated a random variable R following

$$R \sim \text{Bernoulli}(\text{logit}^{-1}(\alpha_R + \beta_R X)), \quad (4)$$

where β_R varied to assess the performance under MCAR and deviations from MCAR to MAR. By varying the strength of the association between X and R , we vary the missingness mechanism assumption from almost MCAR ($\beta = 0.5$) to strong MAR ($\beta = 2.0$). Here, we report only three scenarios with $\beta_R = 2, 1, 0.5$, corresponding to strong, moderate, and weak MAR, respectively. Further, α_r is given values of -1 and -0.2 to study the behavior under 25% and 45% rates (approximately) of missingness on Y . As an example for clarification, note that when α_r is set to -1 with $\beta_R = 0.5$ (to indicate weak MAR), R is simulated from

$\text{Bernoulli}(\text{logit}^{-1}(-1 + 0.5X))$, where $X \sim N(0, 1)$ and $\text{logit}^{-1}(u) = \frac{1}{1 + e^{-u}}$. This setup leads to simulated binary values with an approximate mean value 0.25.

Second, we simulated a nominal categorical variable as a function of X to study the performance of our method with incomplete nominal data. Conditional on X , Y was simulated under a baseline-category logit model

$$\log \left(\frac{P(Y=j)}{P(Y=j^*)} \right) = \beta + \alpha_j x_j \neq j^*,$$

where j^* is the baseline category. We chose $\beta = -1.5$, $\alpha_1 = 1$, and $\alpha_2 = 0.3$ to obtain $P(Y = 1) = 0.09$ and $P(Y = 2) = 0.05$. Further, we considered the same three missingness mechanisms (strong, moderate, and weak MAR as stated in the previous paragraph).

The next stage of the simulation design pertains to the imputation strategy. For imputing the ordinal missing values, we used two methods: a conventional approach using a MVN distribution with naive rounding to nearest observed integer and our calibration approach improving upon this conventional approach. For imputing the nominal values, however, we only used our calibration approach, as the MVN method is clearly not applicable. The R package norm [6] was used in all computations.

These steps of data generation, creation of imputed datasets, and MI estimation of the parameters under a calibration-based method as well as conventional rounding define a single cycle of the simulation experiment. The performance of the calibration-based MI in estimating these unknown quantities is assessed in relation to several criteria presented in Tables II–IV. These criteria are the coverage rate (CR), root mean square error (RMSE), and the average width of the confidence interval (AW). CR is the percentage of times that the true parameter value is covered in the 95% confidence interval. Here the true parameter value is the average parameter estimate across the simulations before the missing values are imposed. In order to see the impact of imposing missing values on the performance of our techniques, we gauge the performance in relation to the estimates based on ‘before deletion’. If a procedure is working well, the actual coverage should be close to the nominal rate of 95% in our study. However, it is important to evaluate the coverage with the other measures because high variances can lead to higher CRs. The performance of the procedure is

regarded to be poor if its coverage drops below 90% [14]. If the procedure results in CRs that are close to 100% or below 85% (too high or too low variances could lead to such CRs), extra caution should be taken when using that procedure. RMSE is an integrated measure of bias and variance, evaluating $\hat{\theta}$ in terms of combined accuracy and precision. RMSE ($\hat{\theta}$) is

defined as $\sqrt{E(\hat{\theta} - \theta)^2}$. Finally, AW is the distance between the average lower and upper confidence interval limits across the confidence intervals obtained in each simulation. A high CR along with narrow, calibrated confidence intervals translates into greater accuracy and higher power.

4.2. Performance in estimating the marginal mean

Under moderate rates of missingness and weak MAR (e.g., see simulation scenarios I and II with $\alpha_R = -1$ and $\beta_R = 0.5$ and $\beta_R = 0.1$), we see estimates with negligible bias on the marginal means. Under MCAR (not reported here), the mean estimator for Y_{mis} is unbiased under MCAR. YHZ derived asymptotic results confirming this for binary data, and our results also confirm this finding for ordinal and nominal data via the present simulation study. Because the relationships imposed by MVN directly inform the predictive distribution of missing data, bias might arise under MAR (e.g., $\beta_R \neq 0$). However, the performance measures including bias show consistency across the MAR scenarios.

Table II reports the performance of the calibration method in estimating simple marginal estimates on means and proportions. The results leading to Table II were obtained from 1000 repeated cycles of our simulation experiment and pertain to the marginal Y distribution with strong skewness for the ordinal variable and non-uniform frequency distribution for the nominal variable. Results (not reported) under more symmetric distributional assumptions show a similar or better performance with the conventional rounding. Under the columns corresponding to 25% missingness, we observe the unbiasedness feature clearly. However, as the rate of missingness increases and a stricter assumption of MAR is introduced, the unbiasedness becomes slightly less obvious. Our calibration-based rounding for the ordinal variable performs significantly better than the MVN-based method rounding to the nearest integer, leading to less bias, better coverage rates, and less RMSE. For the nominal case, as the missingness rate increases, we observe slightly lower coverage rates and slightly increased bias and RMSE AW. Although the performance is still acceptable, a slight caution should be practiced when imputing a nominal variable with high rates of missingness. All the results reported from our simulation experiment are based on 10 imputations (results remained the same under different number of imputations).

4.3. Performance in estimating coefficients

In the next stage of our simulation study, we estimated a hypothetical analyst's model. Such a practice forms a basis to gauge the performance of our method in settings where the incomplete nominal variable is used as a covariate, shedding light into the performance of the method used in the data application. We have two models defining this model (e.g., analyst's model), referred to as M1 and M2, respectively. The first model (M1) imposes a linear regression model on X using Y as ordinal covariate, and the second model (M2) does the same with the nominal covariate Y .

$$X_i = \beta_0^{M_1} + \beta_1^{M_1} Y_i + \varepsilon_i^{M_1}, \quad (5)$$

and for the nominal covariate (referred to as M2):

$$X_i = \beta_0^{M_2} + \beta_1^{M_2} I_{Y_i=1} + \beta_2^{M_2} I_{Y_i=2} + \varepsilon_i^{M_2}, \quad (6)$$

where $\beta_0^{M_1}, \beta_1^{M_1}, \beta_0^{M_2}, \beta_1^{M_2}$, and $\beta_2^{M_2}$ represent the intercept and slope parameters in M1 and M2, and $\varepsilon^{M_i} \sim N(0, \sigma_{M_i}^2)$. As before, the results reported in Tables III and IV pertain to 1000 repeated cycles of our simulation experiment. In each of the cycles, we created 10 imputations, fitted M1 and M2 to each of the imputed dataset and combined the estimates under MI employing the rules by [15]. Whereas we used a total number of 10 imputations, we also repeated the cycles for 20, 30, and 50 imputations, but our results were essentially the same.

Overall, the performance of the calibration method was far better than the current practice of rounding to the nearest integer for ordinal variables (rows corresponding to the $\beta_0^{M_1}, \beta_1^{M_1}$). Across all the simulation scenarios, coverage rates that were between 90% and 95% were attained, whereas conventional rounding led to coverage rates as low as 53%. Small biases were also seen under both approaches as a result of the attenuated relationship under MVN and because of heavily skewed distribution of Y . However, calibration-based rounding outperformed conventional rounding across all scenarios with respect to all the criteria considered here. Our method was further contrasted with the ‘no-rounding’ approach. Horton *et al.* [11] suggested a normal imputation of the binary variable without rounding to obtain an unbiased estimate of the mean. This method performs slightly better than the MVN approach in the ordinal imputation but worse than the calibration approach across all measures of performance.

Results pertaining to the imputation of a nominal variable are very intriguing. These results are seen in Tables III and IV in the rows corresponding to $\beta_0^{M_2}, \beta_1^{M_2}$, and $\beta_2^{M_2}$ each of which represents a regression coefficient on the influence of the underlying Y -category on the response X . Because the use of MVN-based imputation is not possible, only the calibration-based results are given. For moderate rates of missingness, a sound performance of calibration-based rounding is observed with CRs around the nominal rates and low bias and RMSE. As the missingness rates increase to 45%, some attenuation is seen in the estimation of relationship between X and the underlying categories of Y . However, the performance of our method is still acceptable with low biases and more than 91% coverage rates.

5. Application

We now demonstrate our calibration methodology using the NSCSHCN, a cross-sectional survey representative on both the national and state levels, providing a unique opportunity to analyze state policies and unmet need for health care. The NSCSHCN was conducted from April 2000 to October 2002 by the National Center for Health Statistics at the Centers for Disease Control and Prevention using the State and Local Area Integrated Telephone Survey of list-assisted random digit dialing. The Children with Special Health Care Needs Screener, the current standard for identifying CSHCN, asks respondents whether any child in the household needed or used more medical care, mental health, or educational services; therapies or counseling; prescription medicines; or had limitations compared with other children their age because of a medical, behavioral, or other health condition that lasted or was expected to last longer than 12 months. If the household had multiple children who were reported to have special needs, one child was randomly selected. The respondent (over 95% of whom was a parent of a child with special needs) then answered multiple questions about that child’s health, service needs, and utilization. The final national sample included

38,866 CSHCN, with approximately 750 CSHCN in each state and the District of Columbia (except for 1500 CSHCN from Missouri).

Our analysis specifically focused on the determinants of self-reported CSHCN with severe conditions (parents were asked to rank the severity of each child's condition). The correlates of this outcome variable are investigated in relation to two types of factors. The first set comprises the predisposing factors which include age in years, race (white, black, and Other), Hispanic ethnicity and mother's education level (less than high school degree, high school degree, some post-high school, and a 4-year college degree or more). The composition of the respondents with respect to these factors is given in Table I. Nearly 80% of respondents were white, 10.4% were black, and 9.9% were in the other category. About 4.5% of the respondents did not indicate their race. Nearly 9% identified themselves as being Hispanic.

The second set of factors comprises the enabling factors, which include poverty level (100%, 100–199%, 200–299%, 300–399%, 400%+) and child's insurance type (uninsured anytime during the past year, public, both public and private, and private insurance at the time of the survey). Public insurance included Medicaid, State Childrens Health Insurance Information Center (SCHIP), military insurance, Title V or other public insurance. Further, the NSCSHCN included questionnaire items used to make up composite scores on three additional enabling factors: (1) partnership in a decision-making process and satisfaction with the received services of families of CSHCN; (2) CSHCN will receive coordinated ongoing comprehensive care within a medical home; and (3) ease of use of the community-based service systems for the families. The composition of the respondents with respect to the enabling factors was as follows: 22.0% were near poor (9.4% did not report their poverty status), and 11.6% were uninsured; 52.6% received care in a medical home. The majority of respondents had private insurance (59.6%); some had both private and public (11.6%), and 21.9% had only public insurance. Furthermore, 64.4% of the families partnered in decision making, and 61% of the kids received care at home; 79% of the families reported the use of community-based services.

The missingness rates among these items varied from 1% to 10%. If we were to use complete cases only to estimate the correlates of CSHCN and disparities, we would only use 12,412 cases, eliminating around 68% of the sample. As discussed in succeeding paragraphs, discarding this large number of incomplete cases imposes serious threats to the validity of the underlying complete-case-only analyses.

5.1. Multiple imputation inference

We applied our calibration method to create multiple imputations of missing values on the race variable. We also multiply imputed the missing values in the other variables and calibrated the imputations of the dummy variables defining the race. We then analyzed these multiply imputed datasets to study the predictors (and disparities among the minorities) of the likelihood of CSHCN with severe conditions.

We imputed incomplete race values 10 times using the calibration approach and imputed the other variables under MVN. Calibration-based imputation used all variables important to the substantive analysis as well as any information on design and missingness as covariates. Specifically, our imputation model included sample weights and cluster dummies (Primary Sampling Unit (PSU) indicators) to incorporate the sampling design into the imputation model. Our model also included the covariates sex, Hispanic indicator, race, age, mother's education, insurance type, poverty status, family and community partnerships, and condition severity (the binary outcome variable of CSHCN with severe conditions). We included the substantive outcome variable in the imputation model to avoid any inconsistencies of the

imputation model with the post-imputation analysis [16, 17]. We estimated the model coefficients from each of the 10 imputed datasets and combined them using the rules for MI inference for scalar estimands [5]. We summarized these results in Table V. Estimated coefficients are interpreted as means of the estimates from the 10 imputations, and the standard errors combine between-imputation variance (variation due to the missing-data method) with variance associated with sampling (i.e., within imputation variance), conditional on the imputed data. Finally, the estimated rate of missing information allows users to assess the quality of the imputation procedure in terms of its impact on the overall substantive inferences. Because of a well-calibrated imputation procedure, the impact of the missing data on the uncertainty measures remain less than 1%, much less than the rate of raw missingness. Moreover, compared with complete-case-only analyses, MI leads to reduced standard errors highlighting the efficiency of MI over unprincipled case-deletion method (Table V).

Several key points should be made regarding the MI estimates. Blacks and Hispanics are more likely to have CSHCN with severe conditions (33.4% and 8.9%, respectively) than whites. Whereas the effect of the ‘other race’ is significant only at 0.05 level, the complete-case-only analysis reveals that the ‘other race’ and white difference is not significant. Further, sex and mother’s education are correlated significantly with CSHCN with severe conditions, whereas the complete-case analysis finds no significance. However, the complete-case-only analysis concludes no statistically meaningful relationship between sex and CSHCN and borderline significant relationship between education and CSHCN.

6. Discussion

Our primary goal was to improve the current rounding practice often applied within well-established and widely used methods for imputing missing data operating under the MVN assumption. However, the methods of this paper can also be used under any continuous model. Our simulation study indicates that the calibration methodology leads to similar marginal distributions to the observed data. Some of the simulation scenarios suggest the potential bias in estimating relationships under strong MAR with high rates of missingness. When the application has relatively low rates of missingness and/or when the missingness mechanism is weak or moderate MAR, we can safely conclude that the calibration leads to acceptable performance. Although perfectly acceptable results are obtained under a strong MAR assumption, one can further improve the performance by making the imputation model richer by including variables that can potentially explain the missingness [14]. Overall, our findings reveal that the calibration method provides a significant tool for those who analyze incomplete data with categorical and continuous elements but are restricted to use MVN-based MI routines.

As noted by [7], calibration is a very general approach to the validation of an imputation model and can be regarded as a version of the posterior predictive check ([18] pp. 159–172; [19]) commonly used in assessing the fit of Bayesian models. The use of calibration in the context of imputation combines the model-based imputations with an ad hoc rounding step with $g - 1$ free parameters (not included in the imputation model). Thus, its use is a bit restrictive, and negligible biases are unavoidable in some statistics. However, the idea of calibration is very attractive for practitioners with well-established tools such as MVN-based imputation software. A useful form of calibration has been recently applied in other settings, particularly on the diagnostics for MI inference or in Bayesian data analysis [20, 21].

The assumption of auxiliary X variables to be completely observed might be seen as a limitation. However, our objective was to propose and evaluate a simple method of imputing missing nominal or ordinal data using readily available multivariate imputation software for

continuous data with some simple rounding rules. If the ultimate goal is joint imputation, one can follow several alternatives such as an increasingly popular sequential (or variable-by-variable) imputation. Although this approach is criticized for its potential incoherence with the underlying posterior predictive distribution, in most practical survey settings, it performs reasonably well [22, 23]. Our calibration-based work is currently being extended to improve the coherence problems of this sequential imputation techniques.

One potentially very useful extension of calibration is when the incomplete-data problem is further complicated by design features such as clustering. Such complications occur in multi-stage surveys conducted on subjects clustered within naturally occurring groups or longitudinal studies of subjects. Similar to its cross-sectional counterparts, categorical data imputation has often been done using ‘primitive’ rounding under MVN-based routines. As argued above, in some instances with clear deviations from normality, such approximations can be harmful to the inferences. As long as design-specific features are preserved (e.g., varying mean and/or covariance structure across the clusters) in the calibration procedure, such extensions are straightforward using routines implemented in the previous studies [24, 25].

Despite the attractive features of our methods, they should be used cautiously. They are essentially approximate methods based on normal imputation methods and are limited because they have few free parameters that can be used in post-imputation calibration. When the validity of the imputations is critical, as when there are large percentages of missing data, it becomes more worthwhile to impute under joint models for categorical variables or combinations of categorical and continuous variables [6, 26]. Alternatively, sequential or variable-by-variable imputation [22, 23] can accommodate data of miscellaneous types by imputing under a collection of univariate conditional models. However, the potential for inconsistencies due to the absence of a joint model and lack of well-developed tools for convergence can be the major motivation for calibration-based routines.

This paper focused only on a single categorical variable. Our methods can be extended advantageously to multiple categorical variables using the framework of variable-by-variable imputation. Because conditional normality is implied by a joint normal distribution, one can use only the linear regression (i.e., Gaussian-based imputation with covariates) combined with calibration to impute multiple categorical variables without compromising from coherence of the imputation model and yet maintain its practicality in the survey settings with bound, skip patterns.

Acknowledgments

This research was in part supported by the National Center on Minority Health and Health Disparities, National Institutes of Health (grant number P20MD003373).

References

1. McPherson M, Arango P, Fox H, Lauver C, McManus M, Newacheck PW, et al. A new definition of children with special health care needs. *Pediatrics*. 1998; 102(1 Pt 1):137–140. [PubMed: 9714637]
2. Chevarley, FM. Utilization and expenditures for children with special health care needs. Research findings No. 24. Rockville: Agency for Healthcare Research and Quality; 2006.
3. Hill K, Freeman L, Yucel RM, Kuhlthau K. Unmet need among children with special care needs in Massachusetts. *Maternal and Child Health Journal*. 2008; 12(5):650–661. [PubMed: 17899342]
4. Tang M, Hill KS, Yucel RM, Perrin JM, Kuhlthau KA. Medicaid managed care and the unmet need for mental health care among children with special health care needs. *Health Services Research*. 2008; 43(3):882–900. [PubMed: 18454773]

5. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.
6. Schafer, JL. Analysis of Incomplete Multivariate Data. London: Chapman & Hall; 1997.
7. Yucel RM, He Y, Zaslavsky AM. Using calibration to improve rounding in multiple imputation. *The American Statistician*. 2008; 62(2):125–129.
8. Schimert, J.; Schafer, J.; Hesterberg, T.; Fraley, C.; Clarkson, D. Analyzing Data with Missing Values in S-Plus. Seattle, WA: Data Analysis Products Division. Insightful Corporation; 2000.
9. SAS Institute. SAS/Stat Users Guide, Version 8.2. Cary, NC: SAS Publishing; 2001.
10. Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*. 2006; 26:1368–1382. [PubMed: 16810713]
11. Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. *The American Statistician*. 2003; 57:229–232.
12. Demirtas H. Rounding strategies for multiply imputed binary data. *Biometrical Journal*. 2009; 51(4):677–688. [PubMed: 19650057]
13. Little, R.; Rubin, D. Statistical Analysis with Missing Data. 2nd edn.. New York: Wiley; 2002.
14. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive missing-data strategies in modern missing-data procedures. *Psychological Methods*. 2001; 6:330–351. [PubMed: 11778676]
15. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–590.
16. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*. 1994; 10:538–573.
17. Reiter JP, Raghunathan TE, Kinney S. The importance of the sampling design in multiple imputation for missing data. *Survey Methodology*. 2006; 32.2:143–150.
18. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. 2nd edn.. Chapman & Hall; London: 2004.
19. Gelman A, Meng X, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*. 1996; 6:733–807.
20. Gelman A, Mechelen I, Verbeke G, Heitjan D, Meulders M. Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics*. 2005; 61:74–85. [PubMed: 15737080]
21. Abayomi K, Gelman AE, Levy M. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2008; 57:273–291.
22. Raghunathan TE, Lepkowski JM, VanHoewyk J. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*. 2001; 27:1–20.
23. Van Buuren S, Oudshoorn C. Multivariate imputation by chained equations: MICE V1.0 Users Guide. Report PG/VGZ/00.038, TNO Preventie en Gezondheid. 2000
24. Schafer JL, Yucel RM. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of the Computational and Graphical Statistics*. 2002; 11(2):437–457.
25. Carpenter, J.; Kenward, M. Instructions for MLwiN multiple imputation macros. Bristol, UK: Centre for Multilevel Modelling; 2008.
26. Javaras KN, Van Dyk DA. Multiple imputation for incomplete data with semicontinuous variables. *Journal of the American Statistical Association*. 2003; 98:703–715.

Figure 1.
Calibration-based rounding for under-MVN imputation for binary variables.

Table 1

Summary of selected characteristics for all CSHCN with severe conditions among those with complete race values versus those with missing race values. Children with more severe conditions were selected if screened positive for either of the following items: condition affects ability significantly or parent rank of severity 8 or more on scale of 0–10.

CSHN survey item	Observed-race-only %		Cases with missing race %	
	% missing	% 95% CI	% 95% CI	
Severity		20.37 (20.07, 20.49)	25.79 (23.92, 27.65)	
Age (years)	0.07			
0–5		19.41 (18.72, 20.12)	18.60 (15.94, 21.25)	
6–12		47.35 (16.38, 48.24)	49.30 (45.61, 52.44)	
13–17		33.41 (32.54, 34.29)	32.09 (28.81, 35.18)	
Race	4.53			
White		79.64 (78.60, 79.41)		
Black		10.40 (10.70, 10.09)		
Other		9.88 (9.53, 10.13)		
Hispanic	0.35	8.67 (8.53, 8.81)	47.88 (44.98, 50.76)	
Sex (male)	0.07	59.8 (58.9, 60.7)	64.62 (15.93, 21.27)	
Maternal education	3.44			
Less than high school degree		8.37 (8.23, 8.52)	12.81 (10.45, 15.14)	
High school degree		26.75 (26.52, 26.98)	28.57 (25.40, 31.74)	
Some post-high school		25.62 (25.39, 25.84)	28.49 (25.32, 31.65)	
4-year college or more		36.38 (36.13, 36.63)	33.04 (29.70, 36.31)	
Poverty status	9.35			
Below 100%		14.73 (14.54, 14.92)	24.71 (21.40, 28.01)	
100–199%		23.11 (22.88, 23.33)	25.29 (28.01, 21.96)	
200–299%		19.93 (19.71, 20.15)	18.82 (15.83, 21.82)	
300–399%		15.77 (15.57, 15.96)	11.76 (9.29, 14.24)	
400%+		26.46 (26.22, 26.70)	19.41 (16.37, 22.45)	
Insurance status	0.38			
Ever uninsured past year		10.86 (10.70, 11.02)	12.21 (9.96, 14.45)	
Public		21.50 (21.29, 21.71)	27.23 (24.18, 30.28)	
Both public and private		4.80 (4.70, 4.91)	9.86 (7.82, 11.90)	

CSHN survey item	Observed-race-only		Cases with missing race	
	% missing	%	%	95% CI
Private		62.83	50.70	(47.28, 54.13)

CI, confidence interval.

Table II

Summary of the simulation results on calibration-based marginal estimates.

Parameter	25% missingness					45% missingness				
	TV	Est	CR	RMSE	AW	TV	Est	CR	RMSE	AW
Scenario I										
$\mu_{Y^*}^{\text{CALIB}}$	3.4	3.398	0.949	0.019	0.039	3.4	3.391	0.936	0.013	0.041
$\mu_{Y^*}^{\text{NORM}}$	3.4	3.406	0.905	0.063	0.040	3.4	3.449	0.759	0.049	0.043
$\mu_{Y^*}^{\text{NR}}$	3.4	3.402	0.915	0.053	0.040	3.4	3.410	0.819	0.044	0.042
$P(Y_{\text{nom}} = 1)$	0.072	0.073	0.948	0.0024	0.010	0.067	0.061	0.901	0.007	0.009
$P(Y_{\text{nom}} = 2)$	0.051	0.051	0.950	0.0005	0.009	0.051	0.053	0.921	0.002	0.008
Scenario II										
$\mu_{Y^*}^{\text{CALIB}}$	3.4	3.394	0.947	0.101	0.402	3.4	3.389	0.935	0.008	0.040
$\mu_{Y^*}^{\text{NORM}}$	3.4	3.420	0.851	0.186	0.424	3.4	3.416	0.792	0.016	0.042
$\mu_{Y^*}^{\text{NR}}$	3.4	3.408	0.875	0.145	0.421	3.4	3.410	0.839	0.033	0.041
$P(Y_{\text{nom}} = 1)$	0.071	0.072	0.935	0.001	0.010	0.078	0.084	0.910	0.006	0.011
$P(Y_{\text{nom}} = 2)$	0.051	0.051	0.950	0.001	0.008	0.047	0.050	0.929	0.003	0.009
Scenario III										
$\mu_{Y^*}^{\text{CALIB}}$	3.4	3.392	0.922	0.008	0.0401	3.4	3.383	91.3	0.025	0.041
$\mu_{Y^*}^{\text{NORM}}$	3.4	3.445	0.782	0.045	0.0434	3.4	3.494	73.1	0.094	0.047
$\mu_{Y^*}^{\text{NR}}$	3.4	3.414	0.813	0.043	0.0420	3.4	3.449	0.799	0.083	0.044
$P(Y_{\text{nom}} = 1)$	0.072	0.076	0.912	0.004	0.01	0.068	0.070	0.898	0.008	0.01
$P(Y_{\text{nom}} = 2)$	0.047	0.049	0.923	0.002	0.04	0.042	0.049	0.885	0.004	0.03

$\mu_{Y^*}^{\text{CALIB}}$, $\mu_{Y^*}^{\text{NORM}}$, and $\mu_{Y^*}^{\text{NR}}$ are the means of the ordinal variable whose values are imputed under MVN using three different rounding practices: calibration (denoted with the 'CALIB' subscript), rounded to the nearest integer (denoted with the 'NORM' subscript), and no rounding (denoted with the 'NR' subscript).

TV, true value (represents the value before imposing missing data); Est, MI estimate; CR, coverage rate; RMSE, root mean square error; AW, average width. Scenarios I, II, III represent weak MAR ($\beta_T = 0.1$), moderate MAR ($\beta_T = 1$), and strong MAR ($\beta_T = 2$), respectively.

Simulation results assessing the performance of calibration on estimating coefficients for the incomplete ordinal covariate.

Table III

Parameter	25% missingness				45% missingness					
	TV	Est	CR	RMSE	AW	TV	Est	CR	RMSE	AW
Scenario I										
Calibration-based rounding										
$\beta_0^{M_1}$	2.050	2.035	0.939	0.154	0.357	2.049	2.041	0.950	0.022	0.362
$\beta_1^{M_1}$	-0.603	-0.605	0.952	0.042	0.100	-0.603	-0.598	0.954	0.006	0.101
Conventional MVN-based rounding										
$\beta_0^{M_1}$	2.050	1.983	0.916	0.327	0.344	2.049	2.014	0.822	0.043	0.344
$\beta_1^{M_1}$	-0.603	-0.583	0.934	0.086	0.096	-0.603	-0.591	0.817	0.013	0.097
Scenario II										
Calibration-based rounding										
$\beta_0^{M_1}$	2.051	2.070	0.937	0.154	0.412	2.044	2.034	0.952	0.04	0.411
$\beta_1^{M_1}$	-0.603	-0.607	0.957	0.042	0.113	-0.603	-0.596	0.953	0.01	0.113
Conventional MVN-based rounding										
$\beta_0^{M_1}$	2.051	2.024	0.914	0.327	0.335	2.044	1.935	0.704	0.122	0.327
$\beta_1^{M_1}$	-0.603	-0.592	0.921	0.086	0.093	-0.603	-0.559	0.517	0.044	0.089
Scenario III										
Calibration-based rounding										
$\beta_0^{M_1}$	1.638	1.669	0.939	0.154	0.403	1.639	1.681	0.920	0.067	0.463
$\beta_1^{M_1}$	-0.482	-0.470	0.932	0.042	0.113	-0.482	-0.468	0.916	0.017	0.127

Parameter	25% missingness				45% missingness			
	TV	Est	CR	RMSE	TV	Est	CR	RMSE
$\beta_0^{M_1}$	Conventional MVN-based rounding							
	1.638	1.935	0.824	0.326	0.372	1.639	1.554	0.704
$\beta_1^{M_1}$								
	-0.482	-0.56	0.815	0.086	0.102	-0.482	-0.441	0.683

Acronyms and scenarios are the same as before.

Table IV

Simulation results assessing the performance of calibration on estimating coefficients for the incomplete nominal covariate.

Parameter	25% missingness				45% missingness					
	TV	Est	CR	RMSE	AW	TV	Est	CR	RMSE	AW
$\beta_0^{M_2}$	-0.103	-0.099	0.958	0.019	Scenario I					
					0.129	-0.101	-0.093	0.943	0.014	0.131
$\beta_1^{M_2}$	1.111	1.109	0.940	0.182	0.536	1.107	0.995	0.920	0.119	0.583
$\beta_2^{M_2}$	0.529	0.520	0.939	0.204	0.632	0.522	0.485	0.940	0.127	0.717
$\beta_0^{M_2}$	-0.104	-0.096	0.940	0.021	Scenario II					
					0.129	-0.102	-0.09	0.939	0.013	0.131
$\beta_1^{M_2}$	1.108	1.091	0.937	0.182	0.545	1.109	0.988	0.902	0.140	0.574
$\beta_2^{M_2}$	0.527	0.511	0.946	0.210	0.657	0.515	0.431	0.929	0.151	0.694
$\beta_0^{M_2}$	-0.102	-0.096	0.939	0.102	Scenario III					
					0.13	-0.103	-0.089	0.937	0.014	0.131
$\beta_1^{M_2}$	1.104	1.062	0.921	0.237	0.564	1.109	0.934	0.907	0.105	0.574
$\beta_2^{M_2}$	0.528	0.487	0.912	0.296	0.693	0.524	0.475	0.913	0.137	0.693

Acronyms and scenarios are same as before.

Table V

Calibration-based MI estimates of coefficients, standard errors, and per cent missing information from 10 logistic regressions modeling the probability of CSHCN with severe conditions along with estimates using complete-case-only data.

	MI estimates				Complete-case-only			
	Est.	Exp(est.)	SE	p-value	% missing info.	Est.	SE	Exp(est.) p-value
Intercept	-0.429	0.651	0.077	0.000	0.048	-0.555	0.137	0.574 0.000
Sex	-0.091	0.913	0.028	0.001	0.030	-0.050	0.049	0.951 0.307
Poverty level	-0.173	0.841	0.011	0.000	0.034	-0.156	0.020	0.855 < 0.001
Age	0.117	1.124	0.019	0.000	0.023	0.098	0.034	1.102 0.004
Mother's education	-0.088	0.916	0.015	0.000	0.060	-0.055	0.027	0.947 0.046
Time affected	1.300	3.669	0.036	0.000	0.093	1.297	0.062	3.658 < 0.001
Family partners in decision making	-0.403	0.668	0.031	0.000	0.200	-0.373	0.051	0.688 < 0.001
Child receives care in medical home	-0.162	0.851	0.029	0.000	0.067	-0.155	0.050	0.856 0.001
Community-based services organized	-0.376	0.687	0.042	0.000	0.519	-0.460	0.055	0.631 < 0.001
Black	0.288	1.334	0.041	0.000	0.021	0.240	0.076	1.27 0.002
Other race	0.085	1.089	0.043	0.050	0.019	0.016	0.078	1.08 0.843