

**University of Illinois at Chicago  
School of Public Health  
Division of Epidemiology and Biostatistics**

*Technical report#:2007-005  
June 2007*

Title: Rounding strategies for multiply imputed binary data

**Authors: Hakan Demirtas**

**Affiliation: University of Illinois at Chicago, Division of Epidemiology and Biostatistics.**

# Rounding strategies for multiply imputed binary data

Hakan Demirtas\*

April 17, 2008

## Abstract

Multiple imputation has emerged as a frequently used approach in dealing with incomplete data in the last two decades. Gaussian and log-linear imputation models are fairly straightforward to implement for continuous and discrete data, respectively. However, in missing data settings that include a mix of continuous and discrete variables, the lack of flexible models for the joint distribution of different types of variables can make the specification of the imputation model a daunting task. The widespread availability of software packages that are capable of carrying out multiple imputation under the assumption of joint multivariate normality allows applied researchers to address this complication pragmatically by treating the discrete variables as continuous for imputation purposes and subsequently rounding the imputed values to the nearest observed category. In this article, we compare several rounding rules for binary variables based on simulated longitudinal data sets that has been used to illustrate other missing-data techniques. Using a combination of conditional and marginal data generation mechanisms and imputation models, we study the statistical properties of multiple-imputation-based estimates for various population quantities under different rounding rules from bias and coverage standpoints. We conclude that a good rule should be driven by borrowing information from other variables in the system rather than relying on the marginal characteristics and should be relatively insensitive to imputation model specifications that may potentially be incompatible with the observed data. We also urge researchers to consider the applied context and specific nature of the problem, to avoid uncritical and possibly inappropriate use of rounding in imputation models.

**Key Words:** Multiple imputation; Normality; Symmetry; Skewness; Kurtosis

## 1 Introduction

Missing data are ubiquitous in statistical practice. Determining an appropriate analytical strategy in the absence of complete data presents challenges for scientific exploration.

---

\*Hakan Demirtas (e-mail:demirtas@uic.edu) is Assistant Professor of Biostatistics, Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 West Taylor Street, Chicago, IL, 60612.

Missing values can give rise to biased parameter estimates, reduced statistical power, and degraded coverage of interval estimates, and thereby may lead to false inferences (Little and Rubin, 2002).

Advances in computational statistics have produced flexible missing-data procedures with a sound statistical basis. One of these procedures involves multiple imputation (MI), a simulation technique that replaces each missing datum with a set of plausible values. The completed data sets are then analyzed by standard complete-data methods, and the results are combined into a single inferential summary that formally consolidates missing-data uncertainty into the modeling process. The key ideas and benefits of MI are reviewed by Rubin (2004) and Schafer (1997a, 1999a). For an extensive bibliography see Rubin (1996), and for recent reviews see Horton and Kleinman (2007), and Harel and Zhou (2007).

The fundamental step in parametric MI is filling in the missing data by drawing from the conditional distribution of the missing data given the observed data under a proposed model. For continuous data, joint multivariate normality among the variables has often been perceived as a natural assumption, since the conditional distribution of the missing data given the observed data is then also multivariate normal and allows for dependence of missing quantities on observed quantities in linear regression relationships. When all the variables are categorical, a log-linear imputation model can be used (Schafer, 1997a). If the sample size is assumed fixed, the set of cell frequencies in a contingency table has a multinomial distribution, and log-linear models offer a flexible framework for specifying possible dependencies among variables. If there are no restrictions on the parameters other than they are true probabilities, then the model is said to be saturated. With complete data, using a Dirichlet prior distribution for the saturated model leads to a conjugate analysis. The posterior distribution is again Dirichlet with updated parameters involving the data and prior parameters.

Most real data sets consist of a mix of data types, including continuously scaled and discrete variables. Although the general location model with conditional Gaussian structure has been shown to work in some settings (Schafer, 1997a), the joint distribution that leads to the presumably correct imputation model is not always straightforward to formulate. Certain associations may be poorly estimated in many applications if the observed data are sparse. Furthermore, the limitations of conditional Gaussian model have been illustrated by Belin et al. (1999). An alternative is to resort to continuous imputation techniques by treating discrete variables as continuous, and subsequently rounding the imputed values to

the nearest observed category. In this paper, we focus on incomplete binary variables that are regarded as continuous for the purpose of imputation, then rounded in some fashion.

The organization of this article is as follows: In the next section, motivated by a longitudinal data set from psychiatric research, we devise a study where we generate incomplete simulated data sets. Then, we describe a few rounding rules, along with their potential merits and pitfalls. Finally, we implement the rounding strategies under consideration on imputed data sets and evaluate the comparative performances in terms of bias and coverage properties for the population parameters we have chosen. Section 3 includes discussion and concluding remarks.

## 2 A real data-driven simulation study

Describing a real phenomenon by generating an environment within which the process is assumed to operate is not uncommon and is often the only feasible way of evaluation. The idea of creating many imperfect proxies of what is believed to be the truth is predicated upon simulating the performance of a method by proposing a variety of populations and missingness mechanisms capable of producing data like those actually seen; then assessing the behavior of various methods over repeated samples from each population, and identifying methods that seem to perform well for a variety of populations. In this section, we present a simulation study driven by this approach to evaluate the differential performances under competing rounding rules.

Our real-data example that anchors the simulation study comes from Hedeker and Gibbons (1997), who use the data from the National Institute of Mental Health Schizophrenia Collaborative Study. Patients were randomly assigned to receive one of three anti-psychotic medications or a placebo. We collapsed the subjects from the three drug treatments into a single group, because the performance of the three drugs was reported to be quite similar (Hedeker and Gibbons, 1997). The outcome of interest, severity of illness, was measured on an ordinal scale ranging from 1 (normal) to 7 (extremely ill), which we treat as continuous. Of note, there are non-integer values due to multiple raters in the data set. Measurements were planned for weeks 0, 1, 3, and 6, but missing values occurred primarily due to drop-out. A few subjects had missing measurements and subsequently returned; for simplicity we have removed these. (We could have included these cases with non-monotone missingness, as Hedeker and Gibbons (1997) did. We decided to exclude them to simplify

the task of constructing alternative hypothetical population models for our simulations.) The monotone missingness assumption (drop-out) has little or indiscernible bearing on the conclusions drawn in this paper and was merely done for convenience. A small number of measurements were also taken at intermediate time points (weeks 2, 4, and 5) which we also ignore. These exclusions reduced the sample from 1603 subject-observations to 1500. With these exclusions, the sample contains 312 patients who received a drug and 101 who received a placebo. In the drug group, 3 patients dropped out immediately after week 0, 27 dropped out after week 1, 34 dropped out after week 3, and 248 completed the study. In the placebo group, no patients dropped out after week 0, 18 dropped out after week 1, 19 dropped out after week 3, and there were 64 completers. Hedeker and Gibbons (1997) noted that the mean response profiles are approximately linear when plotted against the square root of week, and they express time on the square-root scale in their models. Adopting this convention, we define time to be the square root of week. Mean response profiles for drop-outs and completers in the two groups are shown in Figure 1.

**Figure 1 goes here**

## 2.1 Complete data generation

We generate the complete data via conditional (random-effects) and marginal specifications. The conditional version is based on well-known linear mixed-effects model (Laird and Ware, 1982). Let  $y_i = (y_{i1}, \dots, y_{in_i})^T$  denote the responses for subject  $i$ . The model is

$$y_i = X_i\beta + Z_ib_i + \epsilon_i, \tag{1}$$

where  $X_i$  ( $n_i \times p$ ) and  $Z_i$  ( $n_i \times q$ ) contain covariates,  $\beta$  contains fixed effects,  $b_i \sim N(0, \psi)$  contains random effects, and  $\epsilon_i \sim N(0, \sigma^2 V_i)$ . Times of measurement are often incorporated into  $X_i$  and  $Z_i$ , allowing the response trajectories to vary by subject. Common choices for  $V_i$  include the identity or patterned (e.g. autoregressive or banded) matrices that reflect serial correlation. In this specific example,  $y_i$ 's are the responses for individual  $i$  at weeks 0, 1, 3, and 6. In our simulated population, we assume that  $y_i = X_i\beta + Z_ib_i + \epsilon_i$  where the columns of  $X_i$  are a constant (one);  $G$  (0 for placebo, 1 for drug);  $T$  (square root of week); and  $GT$ . The columns of  $Z_i$  are a constant and  $T$ . The fixed effects are set to  $\beta = (5.36, 0.05, -0.32, -0.65)^T$ , the random effects  $b_i$  are normally distributed with covariance matrix

$$\psi = \begin{bmatrix} 0.35 & 0.04 \\ 0.04 & 0.23 \end{bmatrix},$$

and the elements of  $\epsilon_i$  are independent and normal with variance  $\sigma^2 = 0.60$ .

In the marginal version of the complete data population, we define the mean vector and variance-covariance matrix of the drug and placebo groups separately via normal distribution  $y_i \sim N(\mu, \Sigma)$ , then we stack them. We preserve the relative sample sizes in the original data set. For the drug group,  $\mu = (5.410390, 4.441124, 3.731003, 3.034192)$ , and for the placebo group,  $\mu = (5.360506, 5.037990, 4.804825, 4.574858)$ , with a common variance-covariance matrix

$$\Sigma = \begin{bmatrix} 0.9508797 & 0.3906932 & 0.4204797 & 0.4491759 \\ 0.3906932 & 1.2591041 & 0.8568114 & 1.0507675 \\ 0.4204797 & 0.8568114 & 1.7797993 & 1.4928601 \\ 0.4491759 & 1.0507675 & 1.4928601 & 2.5243826 \end{bmatrix}.$$

After data generation is complete, measurements at the last time point were dichotomized by a cut-off value of 3.5, with 1 corresponding to the state of being sick. The true marginal proportions were 0.4741554 and 0.4741782, for the conditional and marginal data generation specifications, respectively. The true odds ratio values relating the dichotomized endpoint to drug vs. placebo are 0.2103874 and 0.2112482, corresponding to a beneficial drug effect. Obviously, for continuous data under conditional normality assumption, marginal and conditional distribution of responses have the same functional form (normal). For this reason, the true values are very similar. Although it is redundant to use both population mechanisms, we chose to do so anyway for expository purposes. For further explanation, see Section 2.3.

## 2.2 Missing data mechanism

We assume that drop-out occurs by the following selection process: the probability that patient  $i$  drops out immediately *after* week  $w = 0, 1, 3$  is

$$\text{expit}(\alpha_w + \gamma_1 y_{iw} + \gamma_2 y_{iw}^2 + \gamma_3 G),$$

where  $\alpha_0 = -0.69$ ,  $\alpha_1 = 2.27$ ,  $\alpha_3 = 2.48$ ,  $\gamma_1 = -2.02$ ,  $\gamma_2 = 0.24$ , and  $\gamma_3 = -0.87$ . Here,  $\text{expit}(x) = 1/(1 + e^{-x})$ . With this nonresponse mechanism, missing values are assumed to be missing at random (MAR) in the sense defined by Rubin (1976). In the results that

follow, we focus on MAR missingness and leave potential concerns about departures from MAR to the Discussion.

For both the conditional and marginal specifications, simulated trajectories and non-response rates represent a very close match to the real data trends on average. Average percentages of subjects that are available at weeks 0, 1, 3, and 6 in simulated data sets are shown in Table 1. The overall completion rate is about 75% which is fairly typical in a longitudinal setting with four time points.

**Table 1 goes here**

### 2.3 Imputing under different models

We created multiply imputed data sets using 1) *R/Splus* package NORM (Schafer, 1999b) which employs a normal imputation model that imposes multivariate normal distribution on responses with unstructured covariances. 2) *R/Splus* package PAN (Schafer, 1997b) which was developed for imputing multivariate panel data, where a group of variables is measured for individuals at multiple time points. Details are given below:

- *NORM*: Let  $y_{ij}$  denote an individual element of  $Y = (Y_{obs}, Y_{mis})$ ,  $i = 1, 2, \dots, n$ ,  $j = 1, 2, \dots, p$ , where  $Y_{obs}$  and  $Y_{mis}$  stand for the observed and missing portions of the complete data matrix  $Y$ . The  $i^{th}$  row of  $Y$  is  $y_i = (y_{i1}, y_{i2}, \dots, y_{ip})^T$ . Assume that  $y_1, y_2, \dots, y_n$  are independent realizations of a random vector, denoted as  $(Y_1, Y_2, \dots, Y_p)$ , which has a multivariate normal distribution with the mean vector  $\mu$  and covariance matrix  $\Sigma$ ; that is  $y_1, y_2, \dots, y_n | \theta \sim N(\mu, \Sigma)$ , where  $\theta = (\mu, \Sigma)$  is the unknown parameter and  $\Sigma$  is positive definite. When imputations are created under Bayesian arguments, MI has a natural interpretation as an approximate Bayesian inference for the quantities of interest based on the observed data. MI can be performed by first running an EM-type algorithm (Dempster, Laird and Rubin, 1977), and then by employing a data augmentation procedure (Tanner and Wong, 1987), as implemented in some software packages (e.g. SAS procedure PROC MI, *Splus* missing data library). The EM algorithm is useful for two reasons: it provides good starting values for the data augmentation scheme, and it gives us an idea about the convergence behavior. Data augmentation using the Bayesian paradigm has been perceived as a natural tool to create multiply imputed data sets. For further details, see Schafer (1997a) and Schimert et al. (2001). When both  $\mu$  and  $\Sigma$  are unknown, the conjugate class for the multivariate normal data model is the normal inverted-Wishart family. When no

strong prior information is available about  $\theta$ , one may apply Bayes' theorem with the improper prior. In the simulated examples, a noninformative prior was used to reflect a state of relative ignorance, which is often bluntly expressed as "let the data talk". Initial estimates for  $\theta$  are typically obtained by the EM algorithm. Then, a data augmentation scheme is implemented as follows: First, a value of missing data from the conditional predictive distribution of  $Y_{mis}$ ,  $Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta^{(t)})$ , is drawn. Then, conditioning on  $Y_{mis}^{(t+1)}$ , a new value of  $\theta$  from its complete-data posterior,  $\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t+1)})$  is drawn. Repeating these two steps from a starting value  $\theta^{(0)}$  yields a stochastic sequence  $(\theta^{(t)}, Y_{mis}^{(t)}) : t = 1, 2, \dots$  whose stationary distribution is  $P(\theta, Y_{mis}|Y_{obs})$ , and the subsequences  $\theta^{(t)}$  and  $Y_{mis}^{(t)}$  have  $P(\theta|Y_{obs})$  and  $P(Y_{mis}|Y_{obs})$  as their respective distributions. For a reasonably large number of iterations, the convergence to these stationary distributions is achieved. Since the complete-data likelihood is assumed to follow a multivariate normal distribution, drawing from conditional distributions above is relatively straightforward and can be performed by applying sweep operators to subsets of the vector  $\mu$  and the matrix  $\Sigma$ .

- *PAN*: The model used by PAN was designed to preserve the following relationships: (a) Relationships among response variables within an individual at each time point; (b) Growth or change in any response variable within an individual across time points; and (c) Relationships between the response variables and any covariates included in the model. It relies on a multivariate extension of well-known linear mixed-effects models (Laird and Ware, 1982). This type of model separates the fixed effects (commonalities) and the random effects (heterogeneities) which are population-averaged regression coefficients and perturbations due to inter-subject variation, respectively. The computational engine of PAN is a Gibbs sampling algorithm (Gelfand and Smith, 1990) which simulates the unknown quantities in a three-step cycle: (1) Draw subject-specific random effects based on some plausible values assumed for the missing data and the model parameters. (2) Draw new random values of the model parameters based on the assumed values for the missing data and random effects obtained in (1). (3) Draw new random values for the missing data given the values in (1) and (2). Repeating (1), (2) and (3) in turn defines a Markov chain (Gilks, Richardson and Spiegelhalter, 1996). Upon convergence, the final simulated values for the missing data come from the distribution which multiple imputations should be drawn. In the current study, default noninformative prior was used. In addition, fixed and random effects regressor matrices were defined in accordance with the way we simulated complete data. For details of PAN, see Schafer and Yucel (2002) and Schafer (1997b).

Here, NORM is the correct imputation mode for marginally created data in the sense that a multivariate normal distribution is posited for the joint distribution of responses and imputation is conducted with the same assumption. Similarly, PAN is the correct mode for conditionally created data since both data generation and imputation procedure are based on linear mixed effects model that is formed conditionally on random effects. In fact, in terms of data generation, moving from conditional to marginal is straightforward for continuous data under the assumption of normality. The purpose of using both approaches that must be essentially similar is merely illustrative, and is done to see how results change when an incompatible imputation model is incorrectly utilized.

## 2.4 Parameters of interest

Two parameters were considered in the simulations. The odds ratio of treatment group and binary responses, and the marginal mean at the last time point. Other parameters could have been considered. However, we chose to concentrate on the odds ratio and marginal mean since they can signal concerns that one might expect to apply more generally to other target quantities.

## 2.5 Some rounding rules

For what follows, capitalized words in parentheses stand for abbreviations we use in the remainder of this article.

- Crude rounding (*CRUDE*): Rounding the imputed values to 0 or 1 based on the threshold 0.5.
- No rounding (*NOROUND*): Horton, Lipsitz and Parzen (2003) argue that leaving the imputed values as they are may lead to less bias in some cases, although unrounded numbers may be physically implausible. In fact, some parameters of interest such as marginal proportions, regression coefficients when the binary variable is treated as explanatory, and correlation coefficients do not require 0 – 1 categorization, other parameters such as odds ratios do.
- Observed proportions (*OBSPROP*): Rounding could be performed in a way to preserve the observed proportions. Yucel, He and Zaslavsky (2008) suggest a version of this by identifying a cutoff value that leads to the marginal expectations that are

similar to the proportions in the incomplete data. Although this approach has intuitive appeal, when the missingness mechanism wipes away a certain category much more often than the other, the observed proportion might be a misleading quantity to preserve.

- Normal approximation to Binomial (*NORMBIN*): Bernaards, Belin and Schafer (2007) argue that approximating the binomial distribution with multivariate normal is a sensible approach since the imputation model is ordinarily Gaussian. In what they call “adaptive rounding”, threshold is taken to be  $\bar{w} - \Phi^{-1}(\bar{w})\sqrt{\bar{w}(1 - \bar{w})}$ , where  $\bar{w}$  denotes the mean value on a single variable of available binary observations and imputed values, and  $\Phi^{-1}$  is the quantile function of the normal distribution. Imputed values are produced by the multivariate normal imputation procedure, and then rounded based on the above threshold.
- Logistic regression (*LOGIT*): The rounding rules mentioned so far are marginal in the sense that rounding is performed marginally on the incompletely observed variable without regard to associations with the other variables in the system. In an attempt to borrow information from the others, one could run a regression-type model after the crude rounding, as a refinement to figure out how plausible crudely rounded values are. More specifically, a logistic regression model that takes the imputed and rounded variable as the response and other variables as predictors gives the predicted probability that the binary variable equals 1. Then, one can generate a Bernoulli draw where this predicted probability is treated as the probability of drawing a 1. This adjustment corresponds to a situation where we first perform crude rounding, and subsequently we align the rounded versions with respect to how well other variables predict them. In our simulation setting, depression scores at the first three time points and the treatment group were considered as explanatory variables, and the rounded measure at the fourth time point was taken as outcome in the logistic regression equation.
- Refinement via conditional distributions (*RCOND*): In a similar spirit with *LOGIT*, after creating multiply imputed data sets and performing crude rounding, one can calculate the mean and variance-covariance structure of the imputed data and draw the missing elements from a conditional normal distribution (variable of interest given the others) for each subject to assess how reasonably rounded versions are obtained.

In the current set-up, the treatment group was also considered one of the components of the multivariate normal distribution as a procedural step since this variable was used in the data generation process.

## 2.6 Evaluation Criteria

The simulation procedure consisted of complete data generation, imposing missing values, MI under NORM and PAN with data augmentation whose starting values were obtained from the EM algorithm, finding the estimates for the marginal mean and odds ratio, and combining them by Rubin’s (2004) rules. The procedure was repeated 500 times for each of the  $2 \times 2 = 4$  scenarios (two sets of data generation mechanisms, and two imputation models). To make a genuine comparison, identical incomplete data sets were used for NORM and PAN within each scenario for each of the  $N = 500$  replicates in the simulation. The relative performances were evaluated using the following quantities that are frequently regarded as benchmark accuracy and precision measures:

*Standardized bias*: the relative magnitude of the raw bias to the overall uncertainty in the system. If the parameter of interest is  $\theta$ , the standardized bias is  $100 \times \frac{E(\hat{\theta}) - \theta}{SE(\hat{\theta})}$ , where SE stands for standard error. If the standardized bias exceeds 50% in a positive or negative direction, then the bias begins to have a noticeable adverse impact on efficiency, coverage and error rates (Demirtas, 2004).

*Coverage rate*: the percentage of times that the true parameter value is covered in the confidence interval. If a procedure is working well, the actual coverage should be close to the nominal rate (i.e. Type I and Type II error rates are properly controlled). However, it is important to evaluate coverage with the other measures, because high variances can lead to higher coverage rates. We regard the performance of the interval procedure to be poor if its coverage drops below 90% (Collins, Schafer and Kam, 2001).

*Root-mean-square error (RMSE)*: an integrated measure of bias and variance. It is considered to be arguably the best criterion for evaluating  $\hat{\theta}$  in terms of combined accuracy and precision.  $RMSE(\hat{\theta})$  is defined as  $\sqrt{E_{\theta}[(\hat{\theta} - \theta)^2]}$ .

*Average width of confidence interval*: the distance between average lower and upper limits across 500 confidence intervals. A high coverage rate along with narrow, calibrated confidence intervals translates into greater accuracy and higher power.

Under the above specification, *standardized bias* is the pure accuracy measure, *average width* is the pure efficiency measure, *coverage rate* and *RMSE* are the hybrid measures. For

other applications of this evaluation system, see Demirtas (2005a, 2005b, 2007), Demirtas and Hedeker (2007), Demirtas et al. (2007), and Demirtas, Freels and Yucel (2008).

## 2.7 Results

We present findings separately for the marginal mean (Table 2) and the odds ratio relating the dichotomous outcome to drug- vs. -placebo status (Table 3). Table 2 summarizes results for estimating the marginal mean based on six rounding rules that were described in Section 2.5. The length of the complete data vector is 413 as in the original data set. The first two columns represent the four combinations of data generation and imputation models. The number of significant digits varies depending on the quantity of interest. Bold characters call attention to bias and coverage results that do not fall within acceptable limits.

### Table 2 and 3 go here

Close examination of Table 2 reveals the following salient characteristics:

- As mentioned before, conditional and marginal data generation mechanisms lead to the same distributional features. Therefore, the critical aspect is not the data generation, but the imputation model.
- PAN gains a competitive advantage over NORM when multiple response variables are observed or measured, and/or complex associations among responses exist since NORM only allows for simple pairwise correlations. For univariate longitudinal data without complex associations, most rounding rules performed better under NORM in comparison to PAN. Only *LOGIT* and *RCOND* yielded satisfactory performance in terms of adequate coverages and acceptable biases under PAN.
- When imputation is conducted under NORM, all rounding rules except for *OBSPROP* perform well. A limitation of *OBSPROP* is undercoverage due to the underestimation of between imputation variability, which is evident from narrower average widths. When the average estimate is biased, narrow intervals translate into poor coverages. On the other hand, *OBSPROP* has better properties than *CRUDE*, *NOROUND*, and *NORMBIN* under PAN.
- *CRUDE*, *NOROUND*, and *NORMBIN* produce acceptable estimates only when MI model is correctly specified (under NORM). It is interesting that no rounding may also be a plausible option, as argued by Horton, Lipsitz and Parzen (2003).

- From both accuracy and efficiency standpoints, the methods that borrow information from other variables, i.e. *LOGIT* and *RCOND*, appear to deliver better performance.

Table 3 follows the same format as Table 2 except that *NOROUND* was not included, since it would not give rise to an odds ratio. The odds ratio is known to be a more sensitive measure, and results are quite different from the ones in Table 2. Coverage rates seem satisfactory for all five rounding rules; biases do not go beyond the acceptable limits under NORM except for *OBSPROP*. *LOGIT* is the clear winner across all metrics as measured by smaller biases and *RMSE*'s. As before, NORM appears to outperform PAN. *CRUDE* and *NORMBIN* yield better results than *RCOND*.

A natural question arises at this point: Who is going to give us the correct imputation model? What is merely available in practice is incomplete data sets. The results for *LOGIT* appear to support our conjecture that a good rounding rule should somehow incorporate information from other variables. Obviously, given sufficient time, one can find better predictive models that predict how plausible crudely rounded values are, hence lead to a refinement in the rounding process. Our advice to researchers is that they first apply crude rounding, then form a solid model that treats the rounded version as the outcome and other variables as predictors, and enhance the quality of rounding accordingly. If one is interested in parameters that involve multiple incomplete discrete variables, the fundamental “predictive model” concept equally applies; one can develop a prediction model through multivariate versions of logit or probit regression to adjust the rounded versions.

### 3 Discussion

We conclude with the re-iteration of the main conclusion: a good rounding rule should be driven by borrowing information from other variables in the system rather than relying on the marginal characteristics, and should be relatively insensitive to imputation model specifications that may potentially be incompatible with the observed data.

The current work is a comparative study that illustrates competing rounding rules in the imputation context. Some of these rules are not new; however, to the best of our knowledge, the ones that employ a predictive model where crudely rounded numbers are rectified via a predictive regression model, are a novelty. We hope that a comparison of these rules under two different normal-theory imputation methods using simulated longitudinal data sets serves as a building block for more sophisticated situations.

There are some limitations to the present work. First, nonignorable modeling is beyond the scope of this manuscript. The theory of MI does not necessarily require MAR, as MI may also be performed under nonignorable models (Demirtas and Schafer, 2003). But an assumption of MAR is commonly employed for MI. For the purposes of this article, ignorable nonresponse was assumed; our focus was limited to rounding binary variables in a sensible way. Second, other rounding rules can be envisioned; the rules described herein do not constitute an exhaustive list. Third, we do not provide cogent mathematical arguments that would support or discredit any of the rules; rather, our intention was to give pragmatic advice to applied researchers who are accustomed to imputing with flexible and easy-to-implement software (NORM and PAN) that are designed for continuous data, yet who have to deal with incomplete binary variables that are in their data set. Finally, the simulation was limited in scope. However, the simulated scenarios, motivated by a real-world example, should shed some light on relative advantages and drawbacks on these rounding rules. Furthermore, the recommendation that a rounding rule that predicates upon borrowing information from other variables may lead to better properties, is admittedly vague. One can build a broad range of prediction models in an attempt to refine crudely rounded values. The idea of simple rounding and followed by an adjustment through a prediction model is worth exploring further.

## REFERENCES

Belin, T.R., Hu, M.Y., Young, A.S., Grusky, O. (1999). Performance of a general location model with an ignorable missing data assumption in a multivariate mental health services study. *Statistics in Medicine*, 18:3123–3135.

Bernaards, C.A., Belin, T.R., Schafer, J.L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26:1368–1382.

Collins, L.M., Schafer, J.L., Kam, C.H. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6:330–351.

Demirtas, H. (2004). Simulation-driven inferences for multiply imputed longitudinal datasets. *Statistica Neerlandica*, 58:466–482.

Demirtas, H. (2005a). Bayesian analysis of hierarchical pattern-mixture models for clinical trials data with attrition and comparisons to commonly used ad-hoc and model-

based approaches. *Journal of Biopharmaceutical Statistics*, 25:383–402.

Demirtas, H. (2005b). Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 24:2345–2363.

Demirtas, H. (2007). Practical advice on how to impute continuous data when the ultimate interest centers on dichotomized outcomes through pre-specified thresholds. *Communications in Statistics-Simulation and Computation*, 36:871–889.

Demirtas, H., Arguelles, L.M., Chung, H., Hedeker, D. (2007). On the performance of bias-reduction techniques for variance estimation in approximate Bayesian bootstrap imputation. *Computational Statistics and Data Analysis*, 51:4064–4068.

Demirtas, H., Freels, S.A., Yucel, R.M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation*, 78:69–84.

Demirtas, H., Hedeker, D. (2007). Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses. *Statistics in Medicine*, 26:782–799.

Demirtas, H., Schafer, J.L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 22:2253–2575.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, Series B*, 39:1–38.

Gelfand, A.E., Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds). (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London.

Harel, O., Zhou, X.H. (2007). Multiple imputation review of theory implementation and software. *Statistics in Medicine*, 26:3057–3077.

Hedeker, D., Gibbons, R.D. (1997). Application of random effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2:64–78.

Horton, J.H., Kleinman, K.P. (2007). A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61:79–90.

Horton, N.J., Lipsitz, S.R., Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *American Statistician*, 57:229–232.

Laird, N.M., Ware, J.H. (1982). Random-effects models for longitudinal data. *Biomet-*

*rics*, 38:963–974.

Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Second Edition. New York: Wiley.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 21:581–592.

Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91:473–520.

Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley Classic Library.

Schafer, J.L. (1997a). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J.L. (1997b). *PAN: Multiple Imputation for Multivariate Panel Data, Software Library for S-PLUS*. University Park, PA: The Pennsylvania State University, Department of Statistics.

Schafer, J.L. (1999a). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:3–15.

Schafer, J.L. (1999b). *NORM: Multiple Imputation of Incomplete Multivariate Data Under a Normal Model, Software Library for S-PLUS*. University Park, PA: The Pennsylvania State University, Department of Statistics.

Schafer, J.L., Yucel, R.M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11: 437–457.

Schimert, J., Schafer, J.L., Hesterberg, T., Fraley, C., Clarkson, D.B. (2001). *Analyzing Data with Missing Values in S-plus*. Seattle, WA: Data Analysis Products Division, Insightful Corp.

Tanner, M.A., Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association*, 82:528–540.

Yucel, R.M., He, Y., Zaslavsky, A.M. (2008). Using Calibration to Improve Rounding in Imputation. *The American Statistician*, 62:125–129.

Figure 1: Mean observed response in psychiatric trial by treatment group (placebo, drug) and drop-out status (drop-out, completer), plotted versus  $T = \text{square root of week}$ .

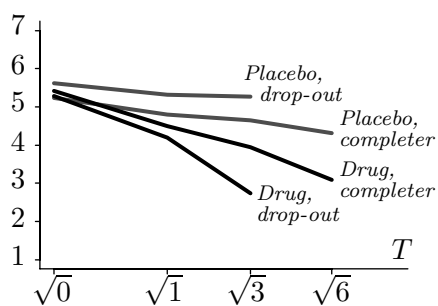


Table 1: The average percentage of available subjects at four measurement weeks for each treatment group.

Treatment/Week	0	1	3	6
Placebo	100	99.5	82.2	63.4
Drug	100	99	90.3	79.5

Table 2: The performance of imputation inferences for the marginal proportions. The number of simulation replicates,  $N$  is 500; the length of the complete data vector,  $n$  is 413 as in the original data.

Data	MI method	Rounding method	Average estimate	Standardized bias	RMSE	Coverage rate	Average width
<i>Conditional</i>	<i>PAN</i>	<i>CRUDE</i>	0.54334	<b>293.35</b>	0.00731	<b>24.2</b>	0.10405
		<i>NOROUND</i>	0.78210	<b>669.95</b>	0.31135	<b>0.0</b>	0.17830
		<i>OBSPROP</i>	0.45690	<b>-63.78</b>	0.03207	<b>83.8</b>	0.09595
		<i>NORMBIN</i>	0.54482	<b>302.89</b>	0.07441	<b>21.6</b>	0.10344
		<i>LOGIT</i>	0.48215	31.67	0.02647	96.6	0.10561
		<i>RCOND</i>	0.47871	18.73	0.02474	97.2	0.10697
<i>Conditional</i>	<i>NORM</i>	<i>CRUDE</i>	0.47352	-2.53	0.02518	97.0	0.10553
		<i>NOROUND</i>	0.47649	8.33	0.02804	95.2	0.10829
		<i>OBSPROP</i>	0.45671	<b>-66.81</b>	0.03234	<b>82.6</b>	0.09634
		<i>NORMBIN</i>	0.47264	-5.76	0.02623	95.8	0.10608
		<i>LOGIT</i>	0.47399	-0.64	0.02549	96.0	0.10348
		<i>RCOND</i>	0.47244	-6.93	0.02484	96.4	0.10510
<i>Marginal</i>	<i>PAN</i>	<i>CRUDE</i>	0.55250	<b>315.75</b>	0.08215	<b>16.0</b>	0.10437
		<i>NOROUND</i>	0.76227	<b>662.97</b>	0.29134	<b>0.0</b>	0.17874
		<i>OBSPROP</i>	0.47319	-3.23	0.02982	<b>89.2</b>	0.09614
		<i>NORMBIN</i>	0.55425	<b>326.22</b>	0.08374	<b>13.4</b>	0.10398
		<i>LOGIT</i>	0.48509	39.97	0.02939	93.8	0.10682
		<i>RCOND</i>	0.48437	38.57	0.02829	95.4	0.10786
<i>Marginal</i>	<i>NORM</i>	<i>CRUDE</i>	0.47542	4.60	0.02696	95.8	0.10595
		<i>NOROUND</i>	0.47485	2.28	0.02949	94.4	0.10789
		<i>OBSPROP</i>	0.47219	-8.19	0.02997	<b>88.8</b>	0.09703
		<i>NORMBIN</i>	0.47444	0.93	0.02809	95.0	0.10652
		<i>LOGIT</i>	0.47520	3.71	0.02762	94.8	0.10384
		<i>RCOND</i>	0.47542	4.67	0.02668	95.4	0.10543

Table 3: The performance of imputation inferences for the odds ratios. The number of simulation replicates,  $N$  is 500; the length of the complete data vector,  $n$  is 413 as in the original data.

Data	MI method	Rounding method	Average estimate	Standardized bias	RMSE	Coverage rate	Average width
<i>Conditional</i>	<i>PAN</i>	<i>CRUDE</i>	0.21135	1.61	0.05963	96.2	0.26008
		<i>OBSPROP</i>	0.27568	<b>102.31</b>	0.09125	96.6	0.32985
		<i>NORMBIN</i>	0.21104	1.09	0.05985	95.0	0.25940
		<i>LOGIT</i>	0.21383	5.34	0.06433	95.2	0.25929
		<i>RCOND</i>	0.24555	<b>53.23</b>	0.07477	96.8	0.29412
<i>Conditional</i>	<i>NORM</i>	<i>CRUDE</i>	0.23486	37.03	0.07042	95.8	0.28272
		<i>OBSPROP</i>	0.25333	<b>63.36</b>	0.08019	96.2	0.29435
		<i>NORMBIN</i>	0.23575	38.19	0.07103	96.0	0.28385
		<i>LOGIT</i>	0.22098	16.28	0.06586	95.4	0.26127
		<i>RCOND</i>	0.24431	49.62	0.07552	96.8	0.28983
<i>Marginal</i>	<i>PAN</i>	<i>CRUDE</i>	0.30254	<b>120.15</b>	0.11872	94.8	0.35114
		<i>OBSPROP</i>	0.32648	<b>162.63</b>	0.13624	94.6	0.36905
		<i>NORMBIN</i>	0.30202	<b>119.30</b>	0.11840	94.8	0.35074
		<i>LOGIT</i>	0.24149	44.07	0.07492	96.2	0.27985
		<i>RCOND</i>	0.25972	<b>71.56</b>	0.08324	96.2	0.30086
<i>Marginal</i>	<i>NORM</i>	<i>CRUDE</i>	0.22677	24.43	0.06533	96.6	0.26218
		<i>OBSPROP</i>	0.22761	27.14	0.06570	96.6	0.26276
		<i>NORMBIN</i>	0.22694	24.73	0.06531	96.8	0.26244
		<i>LOGIT</i>	0.21484	5.76	0.06239	93.8	0.24518
		<i>RCOND</i>	0.23110	31.11	0.06676	97.0	0.26598