Statistics and Its Interface Volume 2 (2009) 449–456

Accuracy versus convenience: A simulation-based comparison of two continuous imputation models for incomplete ordinal longitudinal clinical trials data*

Hakan Demirtas†, Anup Amatya, Oksana Pugach,
John Cursio, Fei Shi, David Morton and Beyza Doganay

Multiple imputation has become an increasingly utilized principled tool in dealing with incomplete data in recent years, and reasons for its popularity are well documented. In this work, we compare the performances of two continuous imputation models via simulated examples that mimic the characteristics of a real data set from psychiatric research. The two imputation approaches under consideration are based on multivariate normality and linear-mixed effects models. Our research goal is oriented towards identifying the relative performances of these methods in the context of continuous as well as ordinalized versions of a clinical trials data set in a longitudinal setting. Our results appear to be only marginally different across these two methods, which motivates our recommendation that practitioners who are not computationally sophisticated enough to utilize more appropriate imputation techniques, may resort to simpler normal imputation method under ignorability when the fraction of missing information is relatively small.

Keywords and phrases: Multiple imputation, Normality, Ignorability, Mixed-effects models, Longitudinal data, Missing data.

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. 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 [1].

Advances in computational statistics have produced flexible missing-data procedures with a sensible statistical basis. One of these procedures involves multiple imputation (MI), a stochastic 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 incorporates missing-data uncertainty into the modeling process. The key ideas and advantages of MI are reviewed by Rubin [2] and Schafer [3, 4]. When a direct maximum likelihood procedure is available for a particular analysis, it may indeed be the convenient method. However, MI still offers some unique advantages for data analysts. First, MI allows researchers to use more conventional models and software; an imputed data set may be analyzed by literally any method that would be suitable if the data were complete. As computing environments and statistical models grow increasingly sophisticated, the value of using familiar methods and software becomes important. Second, there are still many classes of problems for which no direct maximum likelihood procedure is available. Even when such a procedure exists, MI can be more attractive due to fact that the separation of the imputation phase from the analysis phase lends greater flexibility to the entire process. Lastly, MI singles out missing data as a source of random variation distinct from ordinary sampling variability. For an extensive bibliography see [5], for recent reviews see [6] and [7], and for a comparison of MI and likelihood-based methods see [8].

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 this work, we concentrate on incomplete longitudinal data. In addition to employing an imputation model that assumes joint multivariate normality, we use another continuous imputation model that relies on a multivariate extension of well-known linear mixed-effects models [9] for comparison purposes. The second goal of this study is assessing the sensitivity of these imputation approaches to the ordinalized

*The authors thank Don Hedeker for making the data used available on his website.
†Corresponding author.

version of the longitudinal data, where the type of data is clearly unconformable to the underlying modeling assumptions.

The organization of this paper is as follows. In Section 2, we describe the salient operational characteristics of the two imputation models of interest. In Section 3, we describe our simulation philosophy. Subsequently, motivated by a longitudinal data set from psychiatric research, we devise a study where we generate incomplete simulated data sets that resemble the original data trends on average using both ignorable and nonignorable missingness mechanisms. We choose the parameter of interest to be the treatment effect over time, a key quantity in clinical trials, and evaluate the comparative performances of the two imputation models in terms of commonly accepted accuracy and precision measures. A secondary analysis will involve ordinalization of continuous data sets in an attempt to explore the sensitivity of continuous imputation models when applied to ordinal data. Section 4 includes discussion and concluding remarks.

2. IMPUTATION MODELS

We created multiply imputed data sets using 1) *R/Splus* package NORM [10] which employs a normal imputation model that imposes a multivariate normal distribution on responses with unstructured covariances. NORM has the nearly same functionalities as SAS PROC MI which seems to be favored by most practitioners. 2) *R/Splus* package PAN [11] 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 $Y = (Y_{obs}, Y_{mis})$, $i = 1, 2, ..., n$, $j = 1, 2, ..., p$, where i and j index subjects and variables, respectively, and 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 $y_i = (y_{i1}, y_{i2}, ..., y_{ip})^T$. Assume that $y_1, y_2, ..., y_n$ are independent realizations of a random vector, denoted as $(Y_1, Y_2, ..., Y_p)^T$, which has a multivariate normal distribution with the mean vector μ and covariance matrix Σ; that is $y_1, y_2, ..., y_n|\theta \sim N(\mu, \Sigma)$, where $\theta = (\mu, \Sigma)$ is the unknown parameter and Σ 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 Expectation-Maximization (EM)-type algorithm [12], and then by employing a data augmentation procedure [13], 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 [3] and [14]. When both μ and Σ are unknown, the conjugate class for the multivariate normal data model is in the normal inverted-Wishart family. When no strong prior information is available about θ, one may apply Bayes' theorem with an 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 θ 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 θ 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 $(\theta^{(t)}, Y_{mis}^{(t)}) : t = 1, 2, ...$ 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 μ and the matrix Σ.

• *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 [9]. 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 [15] 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 [16]. Upon convergence, the final simulated values for the missing data come from the distribution which multiple imputations should be drawn. In the current study, a 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 [11] and [17].

3. SIMULATION STUDIES

3.1 Philosophy

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. In conjunction with this, creating simulated data sets that are generated around a real data set has been increasingly common in medical statistics, with the rationale being reproducing the real data trends with compatible distributional characteristics. Because there is usually no consensus among statisticians about which of the competing methods is best, many advocate sensitivity analyses that could be performed by trying a variety of methods, or varying the model parameters over a plausible range to see what happens. This approach is valuable, but limited. Instead, we suggest simulating the performance of a method when its assumptions are unmet by proposing a variety of populations that are capable of producing data like those actually seen, simulating behavior of various methods over repeated samples from each population, and subsequently identifying methods that seem to perform well for most of the populations. To elaborate further, suppose we identify a family of models that, from a likelihood standpoint, fit the data equally well. If our basic conclusions about effects of interest do not change drastically over this family, then the scientific validity of these conclusions is enhanced. Conversely, if the answers do exhibit great variation, drawing firm conclusions seems unwise. Robustness of results over the domain of parameters is desirable and fortunate when it occurs. Yet there is another type of analysis which may lead us to prefer one model, M_1, to another, M_2, even when M_1 and M_2 achieve the same likelihood for the current data set. Suppose that we devise a variety of plausible population models, different in nature but all tending to produce samples that resemble the observed data. If, by simulation, we discover that M_1 performs better than M_2 across many of these populations, then we may be more inclined to trust M_1 than M_2 [19, 18]. In this section, we present two simulation studies driven by this philosophy.

3.2 Simulation design

Our real-data example that anchors the simulation study comes from Hedeker and Gibbons [20], 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 [20]. The outcome of interest, severity of illness, was measured on an ordinal scale ranging from 1 (normal) to 7 (extremely ill), which we treated 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 [20] 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. 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 [20] 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. In this trial, the mean profile for the placebo group is slightly declining, indicating mild improvement over time, but the drug group declines more dramatically. Dropout affects the two groups differently. If we classify patients as dropouts or completers, the dropouts in the placebo group appear to be more severely ill than the completers, and show less improvement. In the drug group, however, the opposite occurs: dropouts appear to be less severely ill than completers, and improve more rapidly. One plausible explanation is that those receiving the placebo who experience little or no improvement may be leaving the study to seek treatment elsewhere. On the other hand, those in the drug group who improve dramatically may be dropping out because they feel that treatment is no longer necessary.

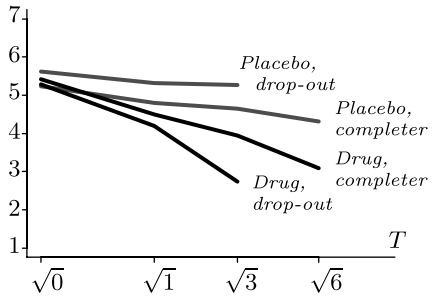

Figure 1. Mean observed response profiles by the treatment group (placebo, drug) and the drop-out status (drop-out, completer), plotted versus T = square root of week.

3.3 Incomplete data generation

We generate the complete data based on well-known linear mixed-effects model [9]. Let $y_i = (y_{i1}, \ldots, y_{in_i})^T$ denote the responses for subject i. The model is

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

where X_i $(n_i \times p)$ and Z_i $(n_i \times q)$ contain covariates, β 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 populations, we assume that $y_i = X_i\beta + Z_i b_i + \epsilon_i$ 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. For the first population, 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 ϵ_i are independent and normal with variance $\sigma^2 = 0.60$.

We assume that dropout occurs by the following selection process: the probability that patient i drops out immediately *before* week $w = 1, 3, 6$, given that he or she has not previously dropped out, is

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

where $\alpha_1 = -1.90$, $\alpha_3 = 0.53$, $\alpha_6 = 0.90$, $\gamma_1 = -1.25$, $\gamma_2 = 0.15$, and $\gamma_3 = -0.90$. For the second population, we set $\beta = (5.36, 0.05, -0.32, -0.65)^T$,

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

and $\sigma^2 = 0.60$; 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 $\gamma_3 = -0.87$. Notice that dropout in the first population is nonignorable, because the probability that y_{iw} is missing depends directly on its value; dropout in the second population is ignorable, depending only on the past observed response. The treatment effects in the two populations are -0.80 and -0.65, respectively. Average response trajectories by treatment group (placebo versus drug) and dropout status (dropout versus completer) for these two populations are shown in Figure 2. As can be seen, the mean response profiles under both nonresponse mechanisms closely mimic the characteristics of the real data trends on average. In addition, the percentage of available subjects (Table 1) and average scores (Table 2) at four measurement weeks among drug and placebo patients for real and simulated data, support this resemblance.

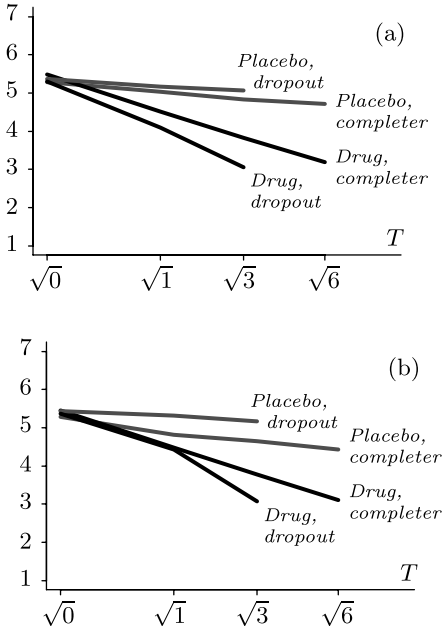

Figure 2. Mean observed response profiles by the treatment group (placebo, drug) and the dropout status (dropout, completer) in simulated data from (a) nonignorable selection population and (b) ignorable selection population.

Table 1. Average percentage of available subjects at four measurement weeks across treatment groups for real data and simulated data under ignorable and nonignorable missingness

	Source	Week0	Week1	Week3	Week6
Overall	Real data	100	99.27	88.38	75.54
	Ignorable	100	99.04	88.41	75.78
	Nonignorable	100	99.17	89.64	74.84
Drug	Real data	100	99.04	90.38	79.49
	Ignorable	100	99.28	91.37	80.50
	Nonignorable	100	99.43	92.12	78.53
Placebo	Real data	100	100	82.18	63.37
	Ignorable	100	98.33	79.29	61.21
	Nonignorable	100	98.35	81.73	62.83

3.4 Parameter of interest

In clinical trials, drug effect over time (drug-time interaction) is a critically important quantity in that the impact of the treatment typically becomes apparent as the time goes by. Since most practitioners regard this interaction effect as a primary measure in assessing the effectiveness of treatments, we treat it as our parameter of interest. The analysis

Table 2. Average severity scores during four measurement weeks for real and simulated data

	Source	Week0	Week1	Week3	Week6
Overall	Real data	5.39	4.57	4.03	3.35
	Ignorable	5.40	4.58	3.94	3.34
	Nonignorable	5.41	4.57	3.94	3.43
Drug	Real data	5.39	4.44	3.80	3.10
	Ignorable	5.41	4.44	3.72	3.07
	Nonignorable	5.44	4.41	3.69	3.13
Placebo	Real data	5.37	4.99	4.79	4.32
	Ignorable	5.36	5.04	4.73	4.40
	Nonignorable	5.32	5.09	4.82	4.57

model for post-imputation data is the linear mixed model (Equation 1) that is presented in Section 3.3.

3.5 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 drug-time interaction, and combining them by Rubin's rules [2]. 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 number of subjects is 413 as in the original data. Under NORM, the model included responses at four measurement weeks as well as the treatment indicator; and under PAN, the population-averaged and subject-specific parts were the same as the one in Section 3.3. Under both imputation models, the number of imputations was 10. The number of EM cycles varied between 100 and 200, therefore following the recommendation by Schafer [3], the number of iterations for the data augmentation procedure was chosen to be 500.

The relative performances were evaluated using the following quantities that are frequently regarded as benchmark accuracy and precision measures:

Standardized bias (SB): the relative magnitude of the raw bias to the overall uncertainty in the system. If the parameter of interest is θ, 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%, then the bias begins to have a noticeable adverse impact on efficiency, coverage and error rates [21].

Coverage rate (CR): 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%.

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 (AW): 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.

Table 3. Performances of NORM and PAN for continuous data under both missingness mechanisms. TV=True value, AE=Average estimate, CR=Coverage rate, SB=Standardized bias, RMSE=Root mean square error, and AW=Average width. The number of subjects is 413, and the number of simulation replicates is 500

	Ignorable		*Nonignorable*	
Quantity	NORM	PAN	NORM	PAN
TV	−0.6504	−0.650	−0.800	−0.800
AE	−0.652	−0.655	−0.714	−0.726
CR	95.1	94.8	81.9	85.4
SB	0.46	3.05	105.07	91.24
RMSE	3.031	3.032	3.059	3.067
AW	0.33	0.32	0.32	0.32

3.6 Results

The results for the two imputation models (NORM and PAN) under the two missingness mechanisms (ignorable and nonignorable) are shown in Table 3. In addition to the quantities mentioned in Section 3.5, we report the true value (TV). The true values were computed based on complete data. Since conditional normality in the linear-mixed effects model implies marginal normality, the true values are the same under NORM and PAN. When the nonresponse mechanism is ignorable, NORM and PAN produced almost identical results in terms of bias, coverage, and efficiency measures (left half of Table 3). These imputation models are designed for ignorable nonresponse, although the theory of MI does not necessarily require ignorability. Imputation inferences can be conducted under nonignorable models as well [18, 19]. As one would expect, in this particular example, the performance of the methods as measured by the key evaluation quantities becomes worse when the missingness mechanism is nonignorable. However, the research goal of this manuscript is not assessing sensitivity for departures from the ignorability assumption. Rather, it is a comparison study between the two well-accepted continuous imputation approaches. It is interesting to note that regardless of their

marginal performances, from a comparative point of view, they fare nearly equally in this example in terms of the bias and efficiency measures we considered.

The fraction of missing information (FMINF) is a major aspect in the imputation world. In this example, FMINF for the regression coefficient of the interaction effect is about 20%. NORM is conceptually, operationally and computationally simpler than PAN, and comparable results suggest that practitioners may prefer NORM over PAN, when FMINF is not too large for the continuous longitudinal data although PAN is specifically designed for this type of data. In other words, PAN would be more correct, accurate and appropriate, however, applied researchers who are not statistically sophisticated enough may rely on NORM on the grounds of simplicity and familiarity with user-friendly software (Windows-based software package is available at *http://www.stat.psu.edu/∼jls/misoftwa.html*).

What if the data type is clearly unconformable with these imputation methods? In the next section, we attempt to answer this question.

3.7 Ordinal data: A sensitivity study for incompatible data types

Here, we investigate a situation where the data are of ordinal type, hence the two imputation models are incompatible with the data.

The original data were ordinal, although there was a minor twist due to multiple raters, as mentioned in Section 3.2. Simulated data sets that we used up to this point were continuous. We proceeded with the same incomplete data generation scheme, then ordinalized the data to seven categories by rounding to the nearest observed category. The original "severity of illness" was scored as 1=normal, not at all ill; 2=borderline mentally ill; 3=mildly ill; 4=moderately ill; 5=markedly ill; 6=severely ill; 7=extremely ill. We also considered a case with four ordinal levels by re-coding the seven ordered categories into four as Hedeker and Gibbons [22] did: (1) normal or borderline mentally ill, (2) mildly or moderately ill, (3) markedly ill, and (4) severely or extremely ill.

After imputation and rounding imputed values to the nearest observed category, we analyzed the resulting data sets by a random intercept and slope mixed-effects ordinal regression model. For subject i at timepoint j, for $c-1$ cumulative logits (here, $c = 4$ or 7), with D denoting *Drug* (0 for placebo, 1 for anti-psychotic drug), and W denoting *Week*, $log\left[\frac{P_{ijc}}{1-P_{ijc}}\right] = \gamma_c - [\beta_0 + \beta_1\sqrt{W_j} + \beta_2 D_i + \beta_3(D_i \times \sqrt{W_j}) + \nu_{0i} + \nu_{1i}\sqrt{W_j}]$, where β's stand for fixed effects, ν_{0i} is the random intercept, ν_{1i} is the random slope, and γ's stand for thresholds ($\gamma_1 = 0$). Random effects are assumed to follow a normal distribution. For the four-category case, in this model, $-\beta_0$ represents the week 0 first logit (category 1 versus 2-4), $\gamma_1 - \beta_0$ the week 0 second logit (1-2 versus 3-4), and $\gamma_2 - \beta_0$ the week 0 third logit (1-3 versus 4) for the placebo group. In terms of the regression parameters, β_1 represents the weekly (in square root units) logit change for placebo patients, β_2 is the difference in the week 0 for drug patients, and β_3 is the difference in the weekly (square root) logit change between drug and placebo groups. The random subject effects ν_{0i} and ν_{1i} represent intercept and slope deviations for subject i, respectively. The seven-category case relies on the same model with different interpretations for the β coefficients. In both the four- and seven-category cases, the parameter of interest is β_3. The simulation setup is identical to the one that is presented in Sections 3.2 to 3.5, except that we have ordinal data rather than continuous data, and the mixed model formulation is non-linear rather than linear at the analysis phase.

Table 4. Performances of NORM and PAN for ordinal data with four and seven levels under both missingness mechanisms. TV=True value, AE=Average estimate, CR=Coverage rate, SB=Standardized bias, RMSE=Root mean square error, and AW=Average width. The number of subjects is 413*, and the number of simulation replicates is* 500

		Ignorable		*Nonignorable*	
Levels	Quantity	NORM	PAN	NORM	PAN
4	TV	−0.859	−0.859	−1.018	−1.018
	AE	−0.826	−0.829	−0.926	−0.935
	CR	94.9	95.2	88.5	90.5
	SB	24.40	22.06	72.40	65.92
	RMSE	0.561	0.563	0.726	0.733
	AW	0.51	0.51	0.51	0.51
7	TV	−0.861	−0.861	−1.035	−1.035
	AE	−0.843	−0.847	−0.950	−0.961
	CR	94.9	94.3	91.6	91.6
	SB	14.92	11.38	69.64	61.26
	RMSE	0.575	0.578	0.745	0.753
	AW	0.50	0.50	0.51	0.50

The results have been shown in Table 4. We have not found substantial differences between NORM and PAN as measured by the key evaluation quantities. Coverage rates, biases, variabilities, and average width of the confidence intervals revealed negligible differences. Furthermore, it is worth noting that the performances of both imputation models are very satisfactory under ignorable nonresponse (we observed a deterioration under nonignorable nonresponse) when the imputation technique is clearly incompatible with the data type. Again, similar results may be due to the fairly low FMINF (about 18%) for β_3. Moreover, we encourage researchers that they use discrete data imputation models [3, 23, 24] when they can, however, they may resort to simpler methods for convenience and simplicity, especially if FMINF is small and if they feel uncomfortable with more appropriate but complicated discrete data methods. We further discuss this issue in Section 4.

Of note, implementing the two imputation models on the real data set yields parameter estimates that are very close

to the true values for the seven-category case we tabulate in Table 4 under ignorable nonresponse (−0.856 for NORM, −0.852 for PAN), as one would expect since the simulated data sets carry the underlying characteristics of the real data on average.

4. DISCUSSION

We did not identify major performance differences between PAN and NORM for both continuous and ordinal versions of data. For continuous longitudinal data, PAN is conceptually more appropriate; for ordinal data, both imputation models are technically unsuitable. However, our limited simulation study demonstrates that imputing with NORM may be a viable approach in some settings given its simplicity and ease of execution from a practitioner's point of view.

There are a few issues that need to be addressed. First, this manuscript is not intended to give definitive advice. The results are based on a limited simulation study. One may argue that simulation studies presented herein are too simplistic compared to many real-life situations where incomplete data structures are more complicated. This argument has certain validity. However, our intention was giving applied researchers some guidance on the relative performances of PAN and NORM for continuous and ordinalized longitudinal mental health data via simulated examples that mimic the features of the real data set. Second, both imputation models are designed to work under ignorable nonresponse, although nonignorable extensions are available [19]. For this reason, they did not perform very well under nonignorable nonresponse. Nevertheless, it is worth to repeat that they yielded similar results. Third, the interaction effect of the treatment group and time was chosen to be the parameter of interest. Other parameters could have been investigated, but drug effect over time is typically regarded as a key quantity in practice. Fourth, for discrete data, we rounded imputed data to the nearest observed category; better rounding rules can be developed. Fifth, we recognize that negligible differences between PAN and NORM may be due to the magnitude of fraction of missing information which was relatively small. Finally and most importantly, whenever feasible, researchers should employ "correct" imputation models such as PAN for continuous longitudinal data, saturated multinomial model [3] or sequential regression imputation [23] for discrete data. Again, if computational capabilities of researchers do not allow them to use these more "appropriate" methods, imputing under multivariate normality assumption may be a reasonable way to proceed in some settings. As always, imputation and analysis modes should be guided by the applied context of the problem as well as convenience versus accuracy considerations.

Received 28 February 2009

REFERENCES

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

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

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

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

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

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

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

[8] Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica* **57** 19–35.

[9] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

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

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

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

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

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

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

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

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

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

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

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

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

[22] Hedeker, D. and Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50** 933–944.

[23] Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* **16** 219–242.

[24] Raghunathan, T. E., Lepkowski, J. M., VanHoewyk, J. and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology* **27** 85–95.

Hakan Demirtas
University of Illinois at Chicago
School of Public Health (MC 923)
1603 W. Taylor Street
Chicago, IL 60612-4336
E-mail address: demirtas@uic.edu

Anup Amatya
University of Illinois at Chicago
School of Public Health (MC 923)
1603 W. Taylor Street
Chicago, IL 60612-4336
E-mail address: aamaty2@uic.edu

Oksana Pugach
University of Illinois at Chicago
School of Public Health (MC 923)
1603 W. Taylor Street
Chicago, IL 60612-4336
E-mail address: opugac1@uic.edu

John Cursio
University of Illinois at Chicago
School of Public Health (MC 923)
1603 W. Taylor Street
Chicago, IL 60612-4336
E-mail address: jcursio@uic.edu

Fei Shi
University of Illinois at Chicago
School of Public Health (MC 923)
1603 W. Taylor Street
Chicago, IL 60612-4336
E-mail address: feishi2@uic.edu

David Morton
University of Illinois at Chicago
School of Public Health (MC 923)
1603 W. Taylor Street
Chicago, IL 60612-4336
E-mail address: dmorto2@uic.edu

Beyza Doganay
Ankara University
Department of Biostatistics
Dekanlik Binasi
Sihhiye, Ankara, Turkey 06100
E-mail address: beyzadoganay@gmail.com