ELSEVIER

Multiple Imputation for Missing Laboratory Data: An Example from Infectious Disease Epidemiology

ZUBER D. MULLA, PHD, BYUNGTAE SEO, PHD, RAMASWAMI KALAMEGHAM, PHD, AND BAHIJ S. NUWAYHID, MD, PHD

PURPOSE: To present multiple imputation (MI) as an appropriate method to address missing values for a laboratory parameter (serum albumin) in an epidemiologic study.
METHODS: A data set of patients who were hospitalized for invasive group A streptococcal infections was accessed. Age was the exposure of interest. The outcome was hospital mortality. Several variables, including serum albumin, were considered to be potential confounders. Of the 201 records, 91 had missing values for serum albumin. The MI procedure in SAS was used to perform 20 imputations of serum albumin by using a Markov chain Monte Carlo approach. Logistic regression was then performed on each of the 20 filled-in data sets, and the results were appropriately combined by using the MIANALYZE procedure.
RESULTS: Age (≥55 years vs. 0–54 years) was not a risk factor for hospital mortality in the complete-case analysis ($n=110$): adjusted odds ratio (OR) = 2.43 (95% confidence interval [CI]: 0.79–7.53). Age was a significant risk factor in the imputed data set ($n=201$): adjusted OR = 3.08 (95% CI: 1.22–7.78).
CONCLUSIONS: Epidemiologists frequently encounter data sets that contain missing values. Traditional missing data techniques such as the complete-subject analysis may lead to biased results. We have demonstrated the use of a novel technique, MI, to account for missing data.
Ann Epidemiol 2009;19:908–914.

KEY WORDS: *Streptococcus pyogenes*, Serum Albumin, Missing Data, Multiple Imputation, Markov Chains, Monte Carlo Methods.

INTRODUCTION

Invasive group A streptococcal (GAS) disease is a condition of major public health importance. Annually, it is estimated that severe GAS disease (including acute rheumatic fever and invasive infections) is responsible for more than 500,000 deaths worldwide (1). One of the severe manifestations of invasive GAS disease is necrotizing fasciitis. In the United States the short-term risk of mortality in patients carrying the diagnosis of GAS necrotizing fasciitis is 24% (2).

In this article we illustrate the use of multiple imputation (MI) to address missing values for a laboratory parameter, serum albumin, in an investigation of predictors of mortality among patients hospitalized for invasive GAS infections. The presence of missing data in epidemiologic research is a common occurrence. A prevalent practice in clinical epidemiology is to simply delete any records that contain missing values for any of the variables under study. This technique is referred to as a complete-subject analysis (3), and it may result in bias. In contrast, imputation methods predict and fill in the missing values by using the observed data (3). MI replaces each missing value with a group of plausible values and then these data sets are analyzed by using standard procedures and then combined in an appropriate manner.

Human serum albumin is a major plasma protein that is primarily synthesized in the liver (4); it has multiple physiological functions. A reduced serum albumin level is usually seen in liver diseases, overhydrated states, starvation, malabsorption, and malnutrition. A decrease in serum albumin may also reflect systemic inflammation in response to insults such as sepsis and trauma (5). Hypoalbuminemia has been linked to increased mortality and morbidity in hospitalized patients and community-dwelling elders (6).

Our previous investigation found that serum albumin at the time of admission, in a univariate model, was strongly associated with hospital mortality in an inverse fashion among individuals with invasive GAS disease (7). This analysis was hampered by a reduction in the sample size

From the Department of Obstetrics and Gynecology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso (Z.D.M., R.K., B.S.N.), and the Department of Epidemiology and Biostatistics, University of South Florida College of Public Health, Tampa (Z.D.M.); the Department of Mathematics and Statistics, Texas Tech University, Lubbock (B.S.); and the Department of Pathology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso (R.K.).

Address correspondence to: Zuber D. Mulla, Department of Obstetrics and Gynecology, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, 4800 Alberta Ave., El Paso, TX 79905. Tel: (915) 545-6710. Fax: (915) 545-6946. E-mail: zuber.mulla@ttuhsc.edu.

Received May 13, 2009; accepted August 9, 2009.

1047-2797/09/$–see front matter
doi:10.1016/j.annepidem.2009.08.002

Selected Abbreviations and Acronyms

CI = confidence interval
EM = expectation-maximization
GAS = group A streptococcal
MAR = missing at random
MCMC = Markov chain Monte Carlo
MI = multiple imputation
OR = odds ratio

from 257 patients to 117 patients due to missing values for serum albumin (7). In our current multivariate analysis, we demonstrate the utility of MI in the setting in which values for this important laboratory parameter are missing for a proportion of the records found in a database of patients with invasive GAS disease.

MATERIAL AND METHODS

Source of Patients and Inclusion Criteria

Demographic and clinical data from an epidemiologic study of invasive GAS disease that was conducted in Florida, the methods of which are described in detail elsewhere, were accessed (8). The database comprises 257 patients who were hospitalized throughout the state of Florida between August 1996 and August 2000 for invasive GAS disease and reported to the Florida Department of Health (Tallahassee).

Invasive GAS disease was defined as isolation of group A *Streptococcus* from a normally sterile site (e.g., blood, cerebrospinal fluid, joint fluid, pleural fluid, or pericardial fluid), and a clinically compatible presentation. The definition of a clinically compatible presentation was one of several entities, including pneumonia, bacteremia in association with cutaneous infection (e.g., cellulitis, erysipelas, or infection of a surgical or nonsurgical wound), deep soft-tissue infection (e.g., myositis), meningitis, peritonitis, osteomyelitis, septic arthritis, postpartum sepsis (i.e., puerperal fever), neonatal sepsis, and nonfocal bacteremia. The original study also included cases of necrotizing fasciitis if GAS was isolated from a nonsterile site.

For each case of invasive GAS disease that was reported by a county health department to the Florida Department of Health, a three-page surveillance case report form was completed after review of the patient's medical record. This chart review was usually performed by an epidemiology staff member at the reporting county health department. Cultures that were positive for GAS (e.g., blood, body fluids, and wounds) and the clinical presentation were recorded on the case report forms. Risk factors for invasive GAS infection, certain laboratory parameters, and antimicrobial therapy that was received during the hospitalization were also noted on the form.

Statistical Analysis

The dichotomous risk factor of interest was the patient's age: 55 years or older versus 0 to 54 years. Age was originally evaluated as a continuous variable. However, exploratory analyses (8) revealed that the risk of hospital mortality did not increase in a linear fashion with age but exhibited dichotomy at age 55. The outcome was hospital mortality. The following variables were considered potential confounders of the association between age and mortality: race, sex, receipt of the antibiotic clindamycin (yes/no), receipt of one or more beta-lactam antibiotics (yes/no), presence of necrotizing fasciitis as noted on the health department surveillance form (present/absent), and serum albumin at the time of admission. For the purposes of the current investigation race was a binary variable: "white" or "other". All of the patients who were of "other" race were "black" except for one individual who was an Asian/Pacific Islander. We defined a beta-lactam antibiotic as any of the penicillins or cephalosporins. Differences in the aforementioned variables by mortality status were tested for statistical significance using the chi-square test for categorical variables and the two-sample *t* test for continuous variables.

The data were analyzed by using the SAS System for Windows Release 9.1.3 (SAS Institute, Inc., Cary, NC). MI was used to address missing values for the serum albumin variable (3, 9–11). We assumed that data were missing at random (MAR) (9, 10, 12). If the probability that a value is missing does not depend on the missing value but does depend on observed quantities (values of variables that were measured), then the missing-data mechanism is called MAR. As an example, assume that a survey captured data on two variables. The two variables are personal income and sex (12). None of the records have missing values for sex. However, a proportion of the records are missing values for personal income. If the probability that a record has a missing value for income does not depend on the value of the income (for example, rich individuals were just as likely as the others to provide their income) but does depend on sex (for example, women were less likely than men to report their income), then the missing data mechanism is called MAR (12).

The majority of the literature on multivariate incomplete data assumes that the data are MAR (9). Assuming M represents the missing-data indicator matrix, X_{obs} represents the observed components of X whereas X_{mis} denotes the missing components, and Φ denotes unknown parameters (9), then MAR can be represented symbolically as

$$f(M|X, \Phi) = f(M|X_{obs}, \Phi) \text{ for all } X_{mis}, \Phi.$$

While it is impossible to test the MAR assumption without additional information, including a larger set of

predictors in the imputation model may make the MAR assumption more plausible (13). The following variables were used in our imputation: mortality status (the outcome), serum albumin level on admission, age, race, sex, necrotizing fasciitis, clindamycin, and beta-lactam.

The Markov chain Monte Carlo method (MCMC) was implemented using the MI procedure (PROC MI) in SAS. MCMC creates multiple imputations by using simulations from a Bayesian prediction distribution (10). In other words, the process involves simulating draws from the posterior distribution of the missing values conditioning on observed values. We assumed our data were from a multivariate normal distribution. Data augmentation applied to Bayesian inference then proceeded by repeating the following two steps (10): the imputation I-step, and the posterior P-step. Using the estimated mean vector and covariance matrix, the I-step simulated the missing serum albumin value for each patient record independently. The P-step simulated the posterior population mean vector and covariance matrix from the complete sample estimates (10). These new estimates were then used by PROC MI in the I-step. Each step depended on the previous step, and these steps were iterated for a sufficient time such that the results were reliable for a multiply imputed data set. A prior distribution was required to obtain the posterior distribution of the parameters in the mean vector and covariance matrix (10). We specified a noninformative prior (Jeffreys') (9).

The MI procedure in SAS was used to perform 20 imputations by invoking the NIMPUTE option (see the Appendix following the references). The default number of imputations for PROC MI is five. The rationale for this increase is as follows: PROC MI is used in conjunction with another procedure, PROC MIANALYZE, which is discussed below. The output generated by PROC MIANALYZE contains a section entitled Multiple Imputation Parameter Estimates. Each of the independent variables is listed in this section with several items of interest including parameter estimates, standard errors, test statistics which are distributed as Student's *t*, and the degrees of freedom associated with each covariate. The reason for the increase in the number of imputations was to ensure that all of the independent variables had at least 100 degrees of freedom associated with them since the *t* distribution approaches the standard normal distribution as the degrees of freedom increases (11). Once the *t* distribution approaches the standard normal distribution, one can be reasonably confident that the standard errors and *p* values are stable and accurate.

The MINIMUM option was set to zero for the serum albumin variable to avoid imputed values that were negative (see Appendix). The initial estimates for MCMC were derived by SAS by using the expectation-maximization (EM) algorithm (10). A detailed description of the EM algorithm is beyond the scope of this article; however, it is a well-known technique for finding maximum likelihood estimates from incomplete data (9). The default number of burn-in iterations when EM estimates are used as starting values is 200 (11). This value was increased to 1000 (see Appendix). The default number of iterations between imputations is 100. This value was increased to 300 (see Appendix). The reason why these two default values were raised was to increase the probability that convergence of the iterative simulations was achieved (11). Unfortunately, there is no definitive test to determine whether convergence was achieved (9). Time-series and autocorrelation plots were inspected for undesirable trends that may have indicated the presence of nonconvergence (11).

Logistic regression was performed on each of the completed datasets using PROC LOGISTIC. Serum albumin was entered as a continuous variable. Since serum albumin was not the main covariate of interest but rather a potential confounder, we did not test the linearity assumption. To clarify, the logistic regression model assumes that the log odds (also known as the "logit") of the outcome increases or decreases in a linear fashion as a continuous covariate increases (14). If a nonlinear association is detected, then typically the continuous variable is converted to a categorical variable (14). However, age (a dichotomous variable), not serum albumin, was our main covariate of interest. Some epidemiologists believe that if a particular independent variable in a logistic regression model is continuous but that it is not the main exposure of interest, then testing the linear risk assumption is not warranted.

The data augmentation method described above assumes that the data set does not contain any missing values for the dependent variable (15), and it assumes that the predictors are continuous and follow a multivariate normal distribution. In contrast, our vector of predictors with the exception of albumin was composed of binary variables. While the aforementioned data augmentation procedure may be used with a mix of categorical and continuous independent variables (11), we chose to delete any records that had missing values for the outcome and any of the binary predictors. Our sample size was reduced from 257 to 201. To clarify, the data set with 201 records had complete data on mortality status, age, race, sex, clindamycin, beta-lactams, and necrotizing fasciitis but did contain records with missing values for serum albumin.

Finally, the MIANALYZE procedure was used to combine the results of the logistic regression models (10). The serum albumin logistic regression point estimates from each of the 20 complete data sets were averaged by SAS using the MIANALYZE procedure. The standard errors associated with the serum albumin point estimates, however, cannot simply be averaged. The reader is referred

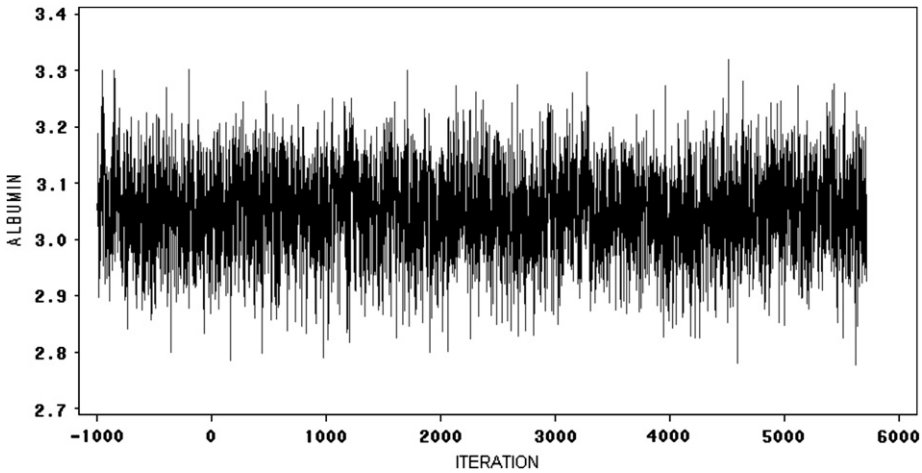

FIGURE 1. Time-series plot of the mean of the serum albumin (in grams per deciliter) covariate by iteration. In healthy individuals serum albumin ranges from 3.5 to 5 g/dL. The X axis indicates there were 1000 burn-in iterations before the first imputation (−1000 to 0). Thereafter, there were 300 iterations between imputations. Twenty imputations were requested. An undesirable pattern (that is, an upward or downward trend) was not observed.

elsewhere (10, 15) for details on the variance estimate associated with the averaged point estimate, but this quantity incorporates both within-imputation variance and between-imputation variance.

With the use of the parameter estimates and standard errors from the PROC MIANALYZE output, odds ratios (ORs) and 95% confidence intervals (CIs) were calculated using standard formulae. For each OR, a Wald chi-square test statistic with one degree of freedom was calculated by dividing the regression coefficient by its standard error and squaring the result and then using SAS' PROBCHI function to calculate a *p* value. We then compared the results of the multiple logistic regression analysis of the imputed data set ($n=201$) with those from the complete-subject data set ($n=110$).

RESULTS

Assessing convergence of the iterative simulations that are conducted in the setting of MCMC is difficult (9). However, inspecting two types of plots (the time-series and the autocorrelation function) may be beneficial in this regard. Fig. 1 displays the time-series plot for the imputed independent variable (serum albumin). The means of serum albumin were plotted against iterations. In our MI procedure, the requested number of imputations was 20. To obtain these, SAS serially generated 7,000 data sets for serum albumin and ignored the first 1,000 to ensure that a "stationary" status was achieved in the Markov chain. This "burn-in" phase is shown on the horizontal axis of Fig. 1 where the iteration ranges from −1000 to 0. The stationary distribution is the distribution of interest (10). When the $(t)-$th generated values for missing serum albumin have the same distribution as that of $(t+s)-$th for any $s = 1,2,3,4,\ldots$, we can conclude that the MCMC sequence is in a stationary status (or converged to the target distribution). It is difficult to completely confirm this, but the data analyst can attempt to verify this by examining the time-series plot of the generated serum albumin or any function of serum albumin, such as the mean of serum albumin. Fig. 1 suggests that a 1,000 burn-in period was quite enough to attain stationary status since an undesirable pattern (a trend up or down) was not observed.

In order to draw independent imputed data sets, for the remaining 6,000 datasets (20 imputations x 300 iterations=6000), we took the 300th, 600th,…., 6,000th data set in order to have 20 independent imputed data sets. The appropriateness of this choice for the number of iterations between imputations can be verified by inspecting the autocorrelation plot for the means of serum albumin (Fig. 2). Fig. 2 indicates that the autocorrelation quickly drops to zero, which implies that a value of 300 iterations was a sufficient amount to result in independent data sets (11).

The following missing data pattern was observed: 54.7% ($n=110$) of the subjects had complete data for serum albumin, whereas 45.3% ($n=91$) were missing serum albumin values. Thirty-five of the 201 patients (17.4%) died in the hospital (Table 1). Patients who expired during their hospital stay were older and had a lower mean serum albumin level on admission than survivors (Table 1).

The distribution of albumin values was not altered by the imputation process. The 20 imputations resulted in 4,020 data points (201 records multiplied by 20 imputations).

AEP Vol. 19, No. 12
December 2009: 908–914

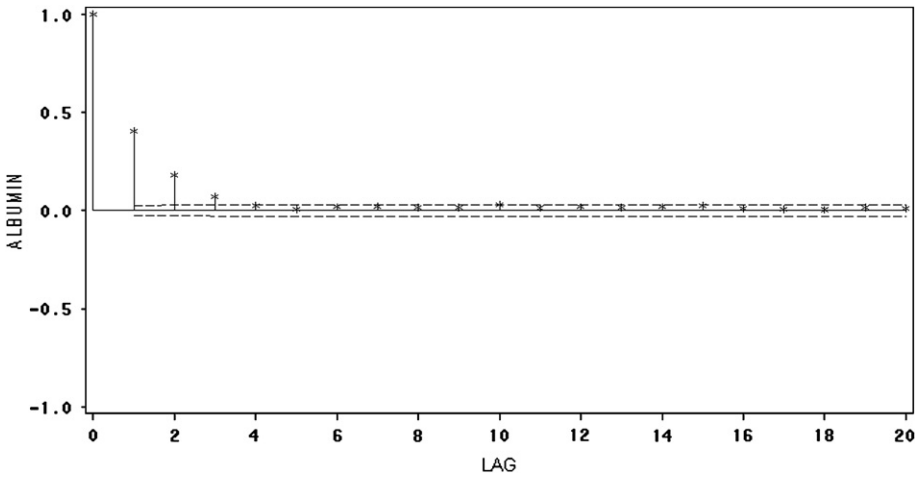

FIGURE 2. Autocorrelations with 95% confidence limits (*dashed lines*) for the mean of the serum albumin covariate at various lags in the sequence of iterations. The goal is to have enough iterations between imputations so that successive values are not correlated (autocorrelation = 0). While the initial values were high, the autocorrelations rapidly decreased to values around zero, indicating an apparent random (and desirable) trend.

The serum albumin (mean ± standard deviation) of the 110 patients in the complete-subject analysis was 3.02 ± 0.82 g/dL, and in the data set with 4,020 patient records, the serum albumin was 3.06 ± 0.83 g/dL. Intensely scrutinizing the imputed/predicted values is not a goal of MI. Rather, the filled-in data sets should be analyzed in an appropriate fashion. In this example, multiple logistic regression was performed on the filled-in data sets. Adjusted ORs for hospital mortality from both the complete-subject data set ($n = 110$) and the imputed data set ($n = 201$) are shown in Table 2. Of the 110 patients, a total of 20 expired during their hospital stay (data not shown). The only statistically significant result from the complete-subject analysis was the OR for serum albumin: adjusted OR = 0.23, 95% CI: 0.10–0.55. This result can be interpreted as follows: for each 1 g/dL increase in serum albumin, there was a 77% reduction in the odds of hospital mortality. Both serum albumin and age were significantly associated with hospital death in the multiply imputed data set. After controlling for the remaining variables shown in Table 2, patients who were 55 years of age or older were 3.08 times as likely as those who were 0 to 54 years of age to die ($p = 0.02$).

TABLE 1. Demographic and clinical characteristics of 201 patients hospitalized for invasive group A streptococcal disease

Characteristic	Died in hospital ($n = 35$)	Survived hospitalization ($n = 166$)	p Value
Demographic variables			
Age ⩾55 yr, *n* (%)	22 (62.9)	69 (41.6)	0.02
White race, *n* (%)	32 (91.4)	128 (77.1)	0.06
Male, *n* (%)	14 (40.0)	84 (50.6)	0.25
Laboratory parameter			
Serum albumin (g/dL),* mean (SD)	2.42 (0.81)	3.15 (0.77)	0.0002
Clinical variables			
Received one or more beta-lactam antibiotics, *n* (%)	27 (77.1)	140 (84.3)	0.30
Received clindamycin, *n* (%)	11 (31.4)	55 (33.1)	0.85
Necrotizing fasciitis, *n* (%)	9 (25.7)	26 (15.7)	0.15

SD = standard deviation.
*Because of missing values, means and SDs are from a sample of 20 decedents and 90 survivors.

DISCUSSION

Analysts frequently encounter data sets that contain missing values (12). Various ad hoc techniques, such as complete-case analysis and mean imputation, have been proposed to addressing missing data. The purpose of this article was to introduce non-statisticians to the sophisticated technique of MI using an MCMC approach in the SAS system.

We found that age was not a statistically significant risk factor for hospital mortality in the complete-subject analysis, but it was a significant predictor of mortality in the completed (imputed) data set. Older age is a known independent risk factor for mortality among individuals with invasive group A streptococcal disease (2). A major advantage of imputation is that it enables the investigator to use all of the available information. This results in an increase in parameter estimation efficiency and the statistical power.

Our MI procedure assumed that all of the variables were normally (Gaussian) distributed. This multivariate normal model is the most popular model for MI (11). While our serum albumin variable followed a normal distribution (data not shown), clearly our categorical variables did not. However, the assumption of multivariate normality is

TABLE 2. Results of the multiple logistic regression analysis of the complete-subject data set and the combined imputed data set: Adjusted odds ratios for the outcome of hospital mortality

	Complete-subject data set $N=110$ patients			Imputed data set $N=201$ patients		
Possible risk factor	Adjusted *OR	95% CI	p Value	Adjusted* OR	95% CI	p Value
Age (≥55 yr vs. <55 yr)	2.43	0.79–7.53	0.12	3.08	1.22–7.78	0.02
White race vs. other race	1.14	0.21–6.13	0.88	2.16	0.57–8.26	0.26
Males vs. females	0.66	0.22–1.99	0.46	0.46	0.19–1.13	0.09
Serum albumin (for every 1 g/dL increase)	0.23	0.10–0.55	0.001	0.23	0.10–0.53	0.001
Treated with clindamycin (yes vs. no)	0.56	0.16–2.02	0.38	0.71	0.25–2.02	0.52
Treated with beta-lactam (yes vs. no)	0.52	0.13–2.13	0.36	0.53	0.17–1.64	0.27
Necrotizing fasciitis (present vs. absent)	1.58	0.44–5.66	0.48	1.17	0.38–3.60	0.78

OR = odds ratio; CI = confidence interval.
*Each odds ratio is adjusted for the remaining variables shown in the table.

harmless with regard to those variables that have no missing data. In our example, only serum albumin was missing at times. Our categorical predictors did not have missing values. Allison (11) writes:

> Typically, there will be some variables with highly skewed distributions and other variables that are strictly categorical. In such cases, is there any value to the normal-based methods…? …there is no problem whatever for variables that have no missing data because nothing is being imputed for them.

Approximately 45% of the patient records in our study were missing values for the serum albumin variable. There is no agreed-upon upper limit for the proportion of records with missing values beyond which a data analyst should not attempt MI. Theoretically, a large missing rate is not a critical problem as long as the assumptions of MAR and multivariate normality are met and the sample size is large because MI takes into account the uncertainty caused by the sizeable missing rate.

Other novel missing data methods such as likelihood-based approaches, weighting methods, and MI using an iterative hot-deck with distance-based donor selection should be explored by epidemiologists (13, 16). Several software packages are available that will allow the investigator to incorporate records into regression models which contain incomplete data. These packages are reviewed by Horton and Kleinman (13).

APPENDIX. SAS PROGRAM USED TO PERFORM MULTIPLE IMPUTATION*

```
/* Perform imputation, create 20 filled-in datasets.
   The dependent variable is hospital mortality and called
outcome. */
   PROC MI DATA=study SEED=9999 OUT=miout
NIMPUTE=20 MINIMUM=0 . . . . . ;
   MCMC NBITER=1000 NITER=300 TIMEPLOT
ACFPLOT;
   VAR albumin outcome age race sex clindamycin
beta_lactam necrofas;
   RUN;
   /* Perform a standard logistic regression analysis on each
of the 20 multiply imputed datasets and output statistics
required in the next phase of analysis, the combination
stage. */
   PROC LOGISTIC DATA=miout OUTEST=outreg
COVOUT DESCENDING;
   MODEL outcome=albumin outcome age race sex
clindamycin beta_lactam necrofas;
   BY _IMPUTATION_;
   RUN;
   /* Combine results in an appropriate fashion and obtain
parameter estimates. */
   PROC MIANALYZE DATA=outreg;
   MODELEFFECTS INTERCEPT albumin outcome age
race sex clindamycin beta_lactam necrofas;
   RUN;
```

* User-defined variables are in lowercase type.

REFERENCES

1. Carapetis JR, Steer AC, Mulholland EK, Weber M. The global burden of group A streptococcal diseases. Lancet Infect Dis. 2005;5:685–694.
2. O'Loughlin RE, Roberson A, Cieslak PR, Lynfield R, Gershman K, Craig A, et al. The epidemiology of invasive group A streptococcal infection and potential vaccine implications: United States, 2000-2004. Clin Infect Dis. 2007;45:853–862.
3. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed Philadelphia: Lippincott Williams & Wilkins; 2008:219.
4. Quinlan GJ, Martin GS, Evans TW. Albumin: biochemical properties and therapeutic potential. Hepatology. 2005;41:1211–1219.
5. Ryan AM, Hearty A, Prichard RS, Cunningham A, Rowley SP, Reynolds JV. Association of hypoalbuminemia on the first postoperative day and complications following esophagectomy. J Gastrointest Surg. 2007;11:1355–1360.
6. Gibbs J, Cull W, Henderson W, Daley J, Hur K, Khuri SF. Preoperative serum albumin level as a predictor of operative mortality and morbidity:

results from the National VA Surgical Risk Study. Arch Surg. 1999;134:36–42.

7. Mulla ZD. Spline regression in clinical research. West Indian Med J. 2007;56:77–79.
8. Mulla ZD, Leaverton PE, Wiersma ST. Invasive group A streptococcal infections in Florida. South Med J. 2003;96:968–973.
9. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed Hoboken (NJ): John Wiley & Sons; 2002.
10. Yuan YC. Multiple Imputation for Missing Data: Concepts and New Development. Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference. Cary (NC): SAS Institute; 2000:1410-1419.
11. Allison PD. Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences 07-136. Thousand Oaks (CA): Sage; 2001.
12. Harrell FE Jr. Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2001:41–42.
13. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat. 2007;61:79–90.
14. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed New York: John Wiley & Sons, Inc; 2000:63–64.
15. Ibrahim JG, Chen M-H, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: a comparative review. J Am Stat Assoc. 2005;100:332–347.
16. Siddique J, Belin TR. Multiple imputation using an iterative hot-deck with distance-based donor selection. Stat Med. 2008;27:83–102.