Appl. Statist. (1996)
45, *No.* 3, *pp.* 335–352

A Multiple-imputation Analysis of a Case–Control Study of the Risk of Primary Cardiac Arrest among Pharmacologically Treated Hypertensives

By TRIVELLORE E. RAGHUNATHAN†

University of Michigan, Ann Arbor, USA

and DAVID S. SISCOVICK

Cardiovascular Health Research Unit, Seattle, USA

[Received November 1993. Final revision August 1995]

SUMMARY

A multiple-imputation method is developed for analysing data from an observational study where some covariate values are not observed. A hybrid approach is presented where the imputations are created under a Bayesian model involving an extended set of variables, although the ultimate analysis may be based on a regression model with a smaller set of variables. The imputations are the random draws from the posterior predictive distribution of the missing values, given the observed values. Gibbs sampling under an extension of the Olkin–Tate general location–scale model is used for the imputation. The method proposed is used to analyse data from a population-based case–control study investigating the association between drug therapy and primary cardiac arrest among pharmacologically treated hypertensives. The sensitivity of the inference to the assumptions about the mechanism for the missing data is explored by creating imputations under several non-ignorable mechanisms for missing data. The sampling properties of the estimates from the hybrid multiple-imputation approach are compared with those based on the complete data and maximum likelihood approaches through simulated data sets. This comparison suggests that much efficiency can be gained through the hybrid approach. Also, the multiple-imputation approach seems to be fairly robust to departures from the assumed normality unless the actual distribution of the continuous covariates is very skew.

Keywords: Gibbs sampling; Logistic regression; Non-ignorable mechanism; Odds ratios; Olkin–Tate model; Sensitivity analysis

1. Introduction

In many observational studies where the relationship between a binary outcome variable such as disease status and an exposure variable is of interest, the logistic regression model is used to eliminate the effect of confounding variables. This can be difficult if the data on the confounding variables are not fully observed. Vach and Blettner (1991) showed that various currently used *ad hoc* methods of correcting for missing values in confounding variables can result in a biased estimation of the regression coefficients of primary interest. In particular, the popular approach, the

†*Address for correspondence*: Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, USA.
E-mail: teraghu@umich.edu

0035–9254/96/45335

so-called complete case analysis, where the inference is based only on the individuals that have fully observed values of the confounding variables, can be biased even under reasonable assumptions about the missing data mechanism. For a 2 × 2 table, Kleinbaum *et al.* (1981) showed that the complete case analysis can lead to unbiased estimates of the odds ratio if the logarithm of the odds of missing either of the two variables satisfies certain additivity properties.

Several researchers have developed methods for fitting a logistic regression model when some covariate values are missing. Little and Schluchter (1985) used the EM algorithm to fit a general location–scale model with missing values. The logistic regression model is a particular case of the general location–scale model considered by them. Dellaportas and Smith (1993) developed a fully Bayesian approach for a generalized linear model by using Gibbs sampling which can be extended to deal with missing values. Vach and Schumacher (1993) developed maximum likelihood, pseudo-maximum-likelihood and probability imputation methods when all the covariates are categorical variables. Robins *et al.* (1994) developed a weighted estimating equation method where the weights are derived from an assumed missing data mechanism. All these developments assume that the data are missing at random (Rubin, 1976).

In this paper we develop a multiple-imputation approach, originally proposed by Rubin (1987) in the context of non-response in sample surveys, for handling missing data with an arbitrary pattern of missing data on both continuous and categorical covariates. Under this scheme we replace each set of missing covariate values by more than one plausible set of values. Each completed data set formed by combining an imputed set of values with the observed data is analysed to obtain estimates and the covariance matrix of the target quantities of interest. These estimates and the covariance matrices are then combined to form a single inference as discussed by Rubin (1987) and Li *et al.* (1991). Though the primary focus of this paper is on fitting logistic regression models, the results are easily extended to other types of model as well.

We develop, like many other applications of the multiple-imputation technique, a hybrid approach in this paper where a Bayesian model is used to create imputations and then a likelihood-based analysis possibly based on a different model is performed on each completed data set. The imputation model used to create imputations usually may involve a larger set of variables than the analysis or the user's ultimate model. The central idea of this approach is to gain efficiency by borrowing strength from the auxiliary variables. For example, suppose that the analyst is interested in fitting a logistic regression model relating a binary variable D with a binary exposure variable E adjusting for a set of confounders X with missing values. The database may contain a set of auxiliary variables Z that may not confound the relationship between D and E but may have predictive power for X. Our approach creates imputations based on the specification of the joint distribution of (D, E, X, Z), although the ultimate analysis of the completed data sets will involve (D, E, X). These imputations may be more efficient as they borrow strength from the auxiliary variables Z and hence the estimate of the regression coefficient for E obtained by using this approach may be more efficient than, for example, the maximum likelihood estimate based on the joint distribution of only (D, E, X). This is indeed true and is demonstrated through simulations in Section 7.

The multiple-imputation approach also lends itself, as we show through an example, to a convenient sensitivity analysis with respect to the assumptions about the missing data mechanism and the probability distribution for observables. For instance, we can create several sets of imputations under a variety of model specifications for either the missing data mechanism or the relevant joint distribution for the observables. We can incorporate the uncertainty due to model misspecifications by combining the inferences under various distributional assumptions.

The rest of the paper is organized in seven sections. Section 2 describes the population-based case–control study that investigated the association between the current use of diuretics for the treatment of hypertension and primary cardiac arrest (PCA). This example motivated the research, the results of which are presented in this paper. Section 3 describes the model assumptions used to create multiple imputations under ignorable missing data mechanisms. Given the model specifications and the observed data, the imputations are the random draws from the posterior predictive distribution of the missing values using Gibbs sampling. Section 4 describes the various steps in Gibbs sampling. Section 5 provides the details on combining inferences from completed data sets. Section 6 describes the application of this procedure to the case–control study described in Section 2. This section also discusses an *ad hoc* modification of the procedure to incorporate information about the missing data mechanism and the model specifications for the observables. Section 7 presents the results from a brief simulation study. Finally, Section 8 is a discussion.

2. Description of Data

Various studies have suggested that the aggressive use of thiazide diuretics to treat hypertension might increase the risk of PCA, also known as sudden cardiac death. PCA is operationally defined as a sudden pulseless condition in the absence of a known non-cardiac condition to account for cardiac arrest. To investigate this further, a population-based case–control study was conducted to examine the association between the current use of diuretics to treat hypertension and the risk of PCA.

The details of the design were described by Siscovick *et al.* (1994). Briefly, the population was defined by those enrolled in the Group Health Cooperative of Puget Sound, a health maintenance organization with a current register of more than 350 000, and who are resident in King County, the largest county in the state of Washington. Each person enrolled has a primary care physician for the diagnosis and treatment of common medical conditions. Group Health maintains a separate medical care record for each person and a computerized pharmacy database that covers all prescriptions filled at any Group Health pharmacy since March 1977.

The cases comprised all incident out-of-hospital PCA as identified by the incident reports of the King County emergency medical services system and the Washington state death tapes during the 14-year period (1977–90) of those enrolled by Group Health at the time of arrest. The controls were a stratified random sample of those who were treated for hypertension as identified by the pharmacy database and medical record review. The stratification variables were age (in decades), gender and calendar year of treatment. Each control was assigned a random index date from the distribution of the event dates of the cases.

The cases and controls who were under 30 or over 79 years of age, or who had a prior history of clinically diagnosed heart disease or any other life threatening conditions such as cancer, liver disease, lung disease and end-stage renal disease as determined by the review of their medical records, were excluded. To be eligible a case (or control) should be exposed to the hypertensive drugs on the event (or index) date. The final tally was 164 cases and 742 controls.

Exposure to hypertensive drugs was ascertained by using the computerized pharmacy database. Since Group Health does not reimburse the cost of prescriptions filled outside a Group Health pharmacy, almost all those enrolled receive their medication from a Group Health pharmacy. Thus, the exposure ascertainment on the event or index date based on the pharmacy database is considered to be accurate. To determine whether an individual was exposed on the event or index date, first a compliance rate of chronic medication use for each individual was estimated; then the estimated compliance rate was applied to the suggested frequency of use for the most recent prescription before the event or index date to determine whether or not the individual had taken the drugs on the event or index date.

To estimate the compliance rate, the following strategy was used. For any two consecutive prescriptions for the hypertensive drugs, the difference between the start date and the end date calculated on the basis of the dose and frequency of the first prescription was expressed as a percentage of the difference in the two prescription dates. These differences were then averaged over the entire time period of the pharmacy history to obtain the overall compliance rate for that individual.

The exposure on the event or index date was divided into six categories: category 1, thiazide-type diuretics alone; category 2, combined thiazide diuretic and a potassium sparing agent; category 3, beta-blocker alone; category 4, thiazide-type diuretics with other non-diuretic hypertensive drugs; category 5, combined thiazide and potassium sparing agent therapy with other non-diuretic drugs; category 6, other non-diuretic drugs. The three comparisons of interest were

(a) category 1 against 2,
(b) category 2 against 3 and
(c) category 4 against 5.

The hypothesis of main interest was that the exposure categories 2 and 5 may provide a protection against PCA compared with exposure categories 1 and 4 respectively. This is of interest as the depletion of potassium is thought to be one of the causal mechanisms for PCA. Comparisons (a) and (c) address this issue. Beta-blockers are another popular regimen to treat hypertension and are usually used as a second line of therapy; hence the comparison was confined to users of a single drug. Since the use of multiple drugs may be an indication of severe hypertension or lack of control of blood pressure, the analysis separated single- and multiple-drug users.

The medical records of these cases and controls were abstracted to ascertain their medical history and other confounding variables. Each subject was enrolled at Group Health for at least a year or had four or more ambulatory care visits to a Group Health clinic. The potential confounding variables were age, gender, years of hypertension, smoking, pretreatment systolic blood pressure, pretreatment diastolic blood pressure, pretreatment pulse, presence of diabetes, any electrocardiograph (ECG) abnormality and three indices from the ECG: the cardiac injury infarction score (CIIS), the left ventricular hypertrophy index (LVHI) and the QT-interval prolongation index QTI (QT-interval prolongation is the distance between the Q-wave and the end of the T-wave in electrocardiograms). Other variables that are less likely to be confounding variables but may be useful for predicting the missing confounding variables include marital status, occupation, number of visits to a Group Health clinic in the previous year, height, weight, the last (before the event or index date) blood pressure and pulse readings, serum potassium, glucose, uric acid and creatinine levels.

However, not all variables were available for every individual. Table 1 provides the mean, standard deviation and percentage with missing values for each variable by case–control status. For each potential confounding variable with missing values, a logistic regression with the missing data indicator as a dependent variable and completely observed variables including the case–control status and the dummy variables for the exposure categories was performed. In several of these models, the regression coefficients corresponding to the exposure status and the interaction between the exposure status and the case–control status were significant thus indicating that the complete case analysis may introduce bias in the assessment of the disease–exposure relationship.

3. Complete Data Model Assumptions

Our strategy is to specify a joint distribution for all the observables and a prior distribution for the parameters. Given this model specification, the imputations are

TABLE 1
Distribution of variables associated with the risk of PCA†

Variable	*Results for cases*			*Results for controls*		
	Mean	*Standard error*	*% missing*	*Mean*	*Standard error*	*% missing*
Age	65	0.7	0	63	0.3	0
Gender (male, %)	58	4	0	53	5	0
Smoking (yes, %)	45	4	17	25	2	13
Pretreatment blood pressure						
Systolic	173	3	32	166	1	30
Diastolic	103	1	32	102	1	30
Pretreatment						
Heart rate	85	1.2	6	81	0.3	7
Diabetes (yes, %)	16	3	0	11	1	0
ECG (abnormal, %)	54	4	25	40	2	27
CIIS	7.4	0.9	30	3.6	0.4	29
LVHI	100	0.9	31	97	0.5	30
QTI	106.4	0.9	31	103.8	0.3	
Years of hypertension	9.5	0.6	9	8.7	0.2	9
Recent blood pressure						
Systolic	144	1.6	0	145	0.7	0
Diastolic	84	1.0	0	86	0.3	0
Glucose	124	4.5	9	111	1.6	8
Creatinine	1.16	0.03	11	1.09	0.02	11
Serum potassium	4.0	0.04	12	4.0	0.02	10
Uric acid	6.7	0.13	10	6.4	0.06	11
No. of visits in previous year	5.4	0.3	0	5.0	0.2	0
Height (in)	67	0.3	0	67	0.1	0
Weight (lb)	173	2.8	0	173	1.7	0

†Number of cases, 164; number of controls, 742.

then created by drawing values from the predictive posterior distribution of the set of missing values given the observed values. For drawing values, we used the Gibbs sampling technique (Gelfand and Smith, 1990). In this section, we describe the complete data model and discuss the Gibbs sampling steps in the next section.

Let U denote a p-dimensional variable that is fully observed on all the individuals in a given random sample of size n. Suppose that the variables that are missing for some individuals consist of an r-dimensional continuous variable Y and a q-dimensional categorical variable Z. The complete data model involves specifying a joint distribution of (Y, Z) given U. A convenient representation of this joint distribution is through the specification of the distribution of Z given U and then the distribution of Y given Z and U.

Suppose that Z_j has C_j levels for $j = 1, 2, \ldots, q$. These categorical variables form a contingency table with $C = \Pi_j C_j$ cells. Let $m = (i_1, i_2, \ldots, i_q)$ denote a cell in the q-way contingency table with C cells. Given a random sample of size n, the C-dimensional vector of cell counts $(n_m = n_{i_1 i_2 \ldots i_q}, m = 1, 2, \ldots, C)$ forms a multinomial random variable with cell probabilities $\pi = (\pi_m = \pi_{i_1 i_2 \ldots i_q}, m = 1, 2, \ldots, C)$. Here we assume that n_m, the number of individuals with $Z_1 = i_1, Z_2 = i_2, \ldots, Z_q = i_q$ where $i_j = 1, 2, \ldots, C_j$, is greater than or equal to 2 in most if not all the cells. Next we specify the conditional distribution of Y given U and Z. Given Z, or equivalently a specific cell m, the continuous responses (or their transforms) Y_{mi}, $i = 1, 2, \ldots, n_m$, are assumed to be identically and independently distributed normal random variables with mean μ_m and the covariance matrix Σ is assumed to be the same across all the C cells.

Even with a few categorical variables, the number of cells C can be very large. The estimation of the cell probabilities may require additional structures such as a log-linear model

$$\log \pi_{i_1 i_2 \ldots i_q} = X_1^{\mathrm{T}} \alpha,$$

where α is an $s_1 \times 1$ vector of regression coefficients and X_1 is an appropriate design matrix that may involve the known values U and log-linear parameters representing the main and the interaction effects.

To reduce the dimensionality of the parameters further, the cell means μ_m are assumed to be normally distributed with mean $V_m \tau$ and covariance matrix Ω where τ is an $s_2 \times 1$ vector of regression coefficients and V_m is an $r \times s_2$ design matrix defined by the categorical variables Z and the fully observed covariates U. This assumption is in tune with an empirical Bayes model that borrows strength from the observations in all the cells to estimate each cell mean μ_m. The fully observed covariates are assumed to have an arbitrary distribution.

The location model without any restriction on the parameters was first proposed by Olkin and Tate (1961) for mixtures of continuous and categorical variables. Little and Schluchter (1985) and Little and Rubin (1987) developed the EM algorithm to estimate the parameters $\{\pi_m, \mu_m, m = 1, 2, \ldots, C, \Sigma\}$ when data on Y or Z are missing for some individuals. They also proposed certain restrictions by using the log-linear and linear regression models to reduce the dimensionality of the parameters. For a similar model, Schafer *et al.* (1993) and Schafer (1994) developed the Gibbs sampling approach to create imputations. The model described in this section imposes additional structure by introducing random effects that allow for borrowing strength across various cells formed by the categorical variables.

Finally the prior distribution for the parameter $\omega = (\alpha, \tau, \Sigma, \Omega)$ is assumed to be of the form

$$\Pr(\omega) \propto |\Sigma|^{-1}|\Omega|^{-(\nu/2+r)} \exp\{\mathrm{tr}(B\Omega^{-1})\},$$

i.e. a flat or non-informative prior for (α, τ, Σ) and a proper inverted Wishart prior for Ω. Technically, a proper prior distribution for Ω ensures a proper posterior distribution for Ω (Raftery and Banfield, 1991; DuMouchel and Waternaux, 1992). The hyperparameters ν (a scalar) and B (a matrix) may be chosen on the basis of external information. By choosing the degrees of freedom ν and the elements of B to be close to 0 (subject to the condition that B is positive semidefinite), the prior distribution can be made diffuse relative to the likelihood.

In the case–control example, the variables with no missing values U consist of 18 variables: age, gender (male $\equiv 1$, female $\equiv 0$), case–control status (case $\equiv 1$, control $\equiv 0$), exposure status (five dummy variables), calendar year, diabetes (yes $\equiv 1$, no $\equiv 0$), compliance, number of visits in the previous year, height, weight, last systolic and diastolic blood pressure and pulse readings and occupation (retired $\equiv 0$, employed $\equiv 1$). The $q = 2$ categorical variables with some missing values Z consist of smoking status (current smoker $\equiv 1$, non- or ex-smoker $\equiv 0$) (Z_1), and ECG status (normal $\equiv 0$, abnormal $\equiv 1$) (Z_2) (thus $C = 4$) and the $r = 11$ continuous variables with some missing values (Y) pretreatment systolic and diastolic blood pressure readings, pretreatment pulse, years of hypertension, cardiac injury infarction score, LVHI, QTI and the serum potassium, glucose, uric acid and creatinine levels.

The predictors (X_1) in the model for the four cell probabilities (smoking status × ECG status) included an intercept term, U, the cross-product of the case–control status with each of the five exposure dummy variables and the cross-product of the case–control status and gender. Thus, the design matrix X_1 contained 25 columns. This choice was the maximum number of regressors that could be used, given the modest sample size ($n = 806$). The design matrix V_m included an intercept term, case–control status, smoking status, ECG status and the two-factor cross-product terms (thus a total of six dummy variables), the cross-product of the variables age, gender, case–control status and exposure status with these six dummy variables and U. Thus V_m had 73 columns. We fixed $\nu = \frac{1}{2}$ and $B = 0.0001I$ where I is an $r \times r$ identity matrix. Thus, our prior distribution is diffuse relative to the likelihood.

4. Gibbs Sampling

Gibbs sampling (Gelfand and Smith, 1990) has received much attention recently. For example, Gelman and Rubin (1992), Gilks and Wild (1992), Raftery and Lewis (1992), Ritter and Tanner (1992) and Smith and Roberts (1993) are a few references in the vast literature that discuss several important computational aspects of this approach.

In the present context, to create M imputations, we draw M values from the joint posterior distribution of the missing set of covariate values and the parameters in the complete data model specification, given the observed data. Briefly, Gibbs sampling involves drawing from each univariate conditional distribution (or that of a subvector) of the missing value or the parameter in a cyclic fashion each time replacing the old values by the most recently drawn values. When the number of cycles tends to

∞, the most recent draw of the set of missing values can actually be considered as a draw from the joint posterior predictive distribution of the missing values given the observed values. In a practical setting, every Pth draw is taken as an approximately independent draw from the joint posterior distribution. It is preferable to ignore the initial few cycles to eliminate the effect of the starting values. Gelman and Rubin (1992) also recommended that several parallel cycles with different starting values be used to assess convergence and also to investigate whether the draws are truly independent. Both Gelman and Rubin (1992) and Raftery and Lewis (1992) suggested approaches for choosing P. Thus, to create M imputed sets of values, we draw MP times from the relevant conditional distributions in the cyclic manner just described.

We now describe the mechanics of Gibbs sampling in the present analysis. Suppose that we have an initial draw of the missing values and the cell means μ_m, $m = 1, 2, \ldots, C$. The initial draw of the set of missing values can be obtained by using simple techniques such as hot deck or random mean imputation (Rubin, 1987). After filling in the set of missing values with the initial draw, the initial draw of the cell means can be obtained by using the mean of a bootstrap sample of observations in each cell.

For notational simplicity, we use the generic notation Rest to denote the values that are being conditioned on, other than the argument of the posterior density. It is straightforward to show that

(a) $\Omega^{-1}|\text{Rest} \sim \text{Wishart}[\{B + \Sigma_m(\mu_m - V_m\tau)(\mu_m - V_m\tau)^{\mathrm{T}}\}^{-1}, C + \nu]$,
(b) $\tau|\text{Rest} \sim \text{normal}(\hat{\tau}, \hat{P})$ where $\hat{P} = (\Sigma_m V_m^{\mathrm{T}}\Omega^{-1}V_m)^{-1}$ and $\hat{\tau} = \hat{P}\Sigma_m V_m^{\mathrm{T}}\Omega^{-1}\mu_m$,
(c) $\Sigma^{-1}|\text{Rest} \sim \text{Wishart}[\{\Sigma_{mi}(Y_{mi} - \mu_m)(Y_{mi} - \mu_m)^{\mathrm{T}}\}^{-1}, N - 2r + 2]$ and
(d) $\mu_m|\text{Rest} \sim \text{normal}(\hat{\mu}_m, S_m)$ where $S_m = (n_m\Sigma^{-1} + \Omega^{-1})^{-1}$ and $\hat{\mu}_m = S_m(n_m\Sigma^{-1}\bar{Y}_m + \Omega^{-1}V_m\tau)$ where $\bar{Y}_m$ is the mean of the n_m Y-values in cell m.

To complete the Gibbs cycle, we need to draw the values of the parameters in the log-linear model and then the missing values, given the drawn values of the parameters. Given the initial draw of the missing categorical variables and U, we can fit a log-linear model by using maximum likelihood. Let $\hat{\alpha}$ denote the maximum likelihood estimate and $\hat{T}$ denote the observed Fisher information matrix. We suggest approximating the posterior distribution of α given Rest by a multivariate normal distribution with mean $\hat{\alpha}$ and covariance matrix $-\hat{T}^{-1}$. This step results in draws only from an approximate posterior distribution of α. Dellaportas and Smith (1993) have discussed an approach to draw from the exact posterior distribution. We chose this strategy, despite its approximate nature, to save computational time.

The final step is to draw the missing values, given the parameters. First we draw the missing categorical outcome variable Z_j for an individual. Given α and the observed values of Z_k, $k \neq j$, the previously drawn values of the missing Z_k, $k \neq j$ and U, we can compute the conditional probability that Z_j takes on a value i_j given Z_k, $k \neq j$ and U for an individual who is missing Z_j where $i_j = 1, 2, \ldots, C_j$. Specifically, this conditional probability is $\phi^{(i_j)}_{i_1,\ldots,i_{j-1},i_{j+1},\ldots,i_q} = \pi_{i_1,\ldots,i_j,\ldots,i_q}/\pi_{i_1,\ldots,*,\ldots,i_q}$ where the asterisk in the subscript denotes summation over the particular margin. The drawn value is then a result of a multinomial experiment with these conditional probabilities. Next, to draw the missing continuous variables, note that given all the categorical variables we then know precisely the cell in which the individual i

belongs. For notational brevity, we shall use obs $\subset \{1, 2, \ldots, p\}$ to denote the indices of the observed continuous variables and $Y_{i,\mathrm{obs}}$ and $Y_{i,\mathrm{mis}}$ to denote the observed and missing continuous variables respectively on individual i. The predictive distribution of $Y_{i,\mathrm{mis}}$ is a multivariate normal with mean

$$\mu_{m,\mathrm{mis}} + \Sigma_{\mathrm{mis,obs}}\Sigma^{-1}_{\mathrm{obs,obs}}(Y_{i,\mathrm{obs}} - \mu_{m,\mathrm{obs}})$$

and covariance

$$\Sigma_{\mathrm{mis,mis}} - \Sigma_{\mathrm{mis,obs}}\Sigma^{-1}_{\mathrm{obs,obs}}\Sigma_{\mathrm{obs,mis}},$$

where $\Sigma_{A,B}$ denotes a submatrix of Σ formed by the row indices in A and column indices in B.

5. Multiple-imputation Inference

This section briefly describes the methods for obtaining inferences from a multiply imputed data set. Suppose that a data analyst wants to fit a logistic regression model with the case–control status as the dependent variable and k independent variables. This model is repeatedly fitted to M completed data sets. Let $\hat{\gamma}_l$ denote the estimate of the regression coefficients γ and $\hat{s}_l$ and $\hat{v}_l$ be the score vector and the observed Fisher information matrix respectively based on the lth completed data. The multiply imputed estimate of γ is $\hat{\gamma} = \Sigma_l\,\hat{\gamma}_l/M$. By the strong law of large numbers,

$$\|\hat{\gamma} - E(\gamma|U, Z_{\mathrm{obs}}, Y_{\mathrm{obs}})\| =$$

$$\left\| \hat{\gamma} - \frac{\int \gamma\, L(\gamma|U, Z, Y)\,\Pr(Z_{\mathrm{mis}}, Y_{\mathrm{mis}}|U, Y_{\mathrm{obs}}, Z_{\mathrm{obs}})\,\mathrm{d}\gamma\,\mathrm{d}Y_{\mathrm{mis}}\,\mathrm{d}Z_{\mathrm{mis}}}{\int L(\gamma|U, Z, Y)\,\Pr(Z_{\mathrm{mis}}, Y_{\mathrm{mis}}|U, Y_{\mathrm{obs}}, Z_{\mathrm{obs}})\,\mathrm{d}\gamma\,\mathrm{d}Y_{\mathrm{mis}}\,\mathrm{d}Z_{\mathrm{mis}}} \right\|$$

converges to 0 as both the number of imputations M and the sample size n tend to ∞, where $L(\gamma|U, Z, Y)$ is the complete data logistic model likelihood, Y_{obs}, Z_{obs}, Y_{mis} and Z_{mis} denote the observed and missing components of the data and $\Pr(Y_{\mathrm{mis}}, Z_{\mathrm{mis}}|U, Y_{\mathrm{obs}}, Z_{\mathrm{obs}})$ is the joint posterior density of the missing values under the model assumptions stated in Section 3. Thus the multiply imputed estimate can be viewed as an approximate Bayes estimate with a flat prior for γ and the predictive distribution for the missing values given the observed values as specified in Section 3.

To obtain the information matrix based on the observed data (U, Y_{obs}, Z_{obs}), we can use the representation of the observed data information matrix in terms of the complete data score vector and information matrix given by Louis (1982),

$$I_{\mathrm{obs}} = E(B|U, Y_{\mathrm{obs}}, Z_{\mathrm{obs}}) - E(SS^{\mathrm{T}}|U, Y_{\mathrm{obs}}, Z_{\mathrm{obs}})$$

Where $B = \partial^2\{\log L(\gamma|U, Y, Z)\}/\partial\gamma\,\partial\gamma^{\mathrm{T}}$ and $S = \partial\{\log L(\gamma|U, Y, Z)\}/\partial\gamma$ are the complete data information matrix and the score vector based on the particular logistic model. Again, by the strong law of large numbers, $\bar{v} = \Sigma_l\,\hat{v}_l/M$ and $\bar{S} = \Sigma_l\,\hat{s}_l\hat{s}_l^{\mathrm{T}}/M$

approximate the first and second term on the right-hand side respectively. Thus the approximate asymptotic covariance matrix of the multiple-imputation estimate is $\hat{V} = (\bar{v} - \bar{s})^{-1}$. An alternative approximation of the covariance matrix is given by

$$\hat{V}' = \sum_{l} \hat{v}_l^{-1}/M + (1 + M^{-1}) \sum_{l} (\hat{\gamma}_l - \hat{\gamma})(\hat{\gamma}_l - \hat{\gamma})^{\mathrm{T}}/(M - 1).$$

If $M < k$, then the second term on the right-hand side may not be positive semidefinite in which case the covariance matrix is taken to be

$$\hat{V}'' = (1 + r_M) \sum_{l} \hat{v}_l^{-1}/M,$$

where $r_M = (1 + M^{-1})\, \Sigma_l\, d_l/k(1 - M^{-1})$ and $d_l = (\hat{\gamma}_l - \hat{\gamma})^{\mathrm{T}}\, \hat{v}_l(\hat{\gamma}_l - \hat{\gamma})$.

For testing the hypothesis H_0: $A\gamma = \eta_0$ where A is a $w \times k$ matrix of known constants, Wald's test statistic

$$D_M = (A\hat{\gamma} - \eta_0)^{\mathrm{T}}(A\hat{V}''L^{\mathrm{T}})^{-1}(A\hat{\gamma} - \eta_0)/w$$

is referred to an F-distribution on w and $\nu = 4 + (t - 4)\{1 + (1 - 2t^{-1})/r_M\}^2$ degrees of freedom where $t = w(M - 1)$. If $t \leqslant 4$, then the number of degrees of freedom for the denominator is $\nu' = \frac{1}{2}(M - 1)(w + 1)(1 + 1/r_M)^2$. For a justification see Rubin (1987), Rubin and Schenker (1986) and Li *et al.* (1991).

6. Results from Data Analysis

In the case–control example, 100 imputations were obtained, taking every 200th draw in each of the 10 parallel Gibbs sequences. The choice of 200 was made on the basis of the convergence statistic $\hat{R}$ defined by Gelman and Rubin (1992) (p. 461) for each parameter or missing value. After creating multiply imputed data sets, a logistic regression model was fitted with the case–control status as a dependent variable, and the exposure status, pretreatment systolic blood pressure and pulse, smoking, ECG abnormality, diabetes, age, gender and years of hypertension as independent variables. The adjusted odds ratios for exposure categories 1 and 3 using the exposure category 2 as the reference group (for comparisons (a) and (b)) and for comparison (c) the adjusted odds ratio for exposure category 4 relative to 5 were obtained. The confidence intervals for the adjusted odds ratios were obtained by using the normal reference distribution and $\hat{V}$ as the approximate covariance matrix of the regression coefficients. The other two covariance matrices ($\hat{V}'$ or $\hat{V}''$) or the t reference distribution resulted in the same interval estimates up to two decimal places.

The results are summarized in Table 2. The top entry in each group is the lower 95% confidence limit, the middle entry is the adjusted odds ratio and the bottom entry is the upper 95% confidence limit. The first set of three rows is based on the complete case analysis, whereas the remaining sets provide the results for various values of the number of imputations. The complete case analysis differs somewhat from the multiple-imputation analysis. On the basis of the complete case analysis, we would conclude that the risk associated with a beta-blocker is similar to the risk based on combined therapy, whereas all the multiple-imputation analyses suggest

TABLE 2
Adjusted odds ratios (and their 95% confidence limits) comparing different regimens†

Method	*Odds ratios for the following comparisons:*		
	Risk of thiazide compared with combined therapy (single drug)	*Risk of beta-blocker compared with combined therapy (single drug)*	*Risk of thiazide compared with combined therapy (multiple drug)*
Complete case analysis (cases, 66; controls, 316)	1.38	0.54	0.64
	3.91	2.01	2.22
	10.94	7.34	7.84
Multiple imputation ($M = 5$)	1.21	0.92	0.86
	3.22	2.32	2.95
	12.24	9.33	8.97
Multiple imputation ($M = 15$)	1.34	0.98	1.11
	3.22	2.95	3.01
	7.23	8.12	7.57
Multiple imputation ($M = 50$)	1.32	0.98	1.12
	3.23	2.99	3.01
	7.21	7.97	7.12
Multiple imputation ($M = 100$)	1.33	0.98	1.11
	3.21	2.99	3.01
	7.22	8.11	7.11

†Number of cases, 164; number of controls, 742; the top entry of each group is the lower 95% confidence limit, the middle entry is the adjusted odds ratio and the bottom entry is the upper 95% confidence limit.

that there may be a modest risk associated with beta-blockers compared with the combined therapy. Similarly, the complete case analysis also underestimates, though to a lesser extent, the risk of PCA when using thiazide with supplemental non-diuretic drugs compared with the combined therapy with supplemental non-diuretic drugs. All the analyses show that people on thiazide alone have a significantly higher risk of PCA compared with the combined therapy alone. The multiply imputed confidence intervals are much shorter than the complete case analysis confidence intervals. This is hardly a surprise, given that 40% of the cases and 43% of the controls were deleted from the analysis. Also, the multiple imputations with $M = 15$, $M = 50$ and $M = 100$ give similar results.

A sensitivity analysis was also conducted by creating imputations under certain non-ignorable missing data mechanisms. While imputing the smoking status, it was assumed that the smokers are less likely to have their smoking status missing and an individual with a missing ECG value would be more likely to be normal. Dropping the subscripts for brevity, conditionally on the covariate, let ϕ_1 denote the probability of being a smoker under an ignorable model. For imputing the smoking status, $\theta_1 = \delta_1\phi_1$ was used as the probability of being a smoker where $\delta_1 < 1$. For ECG status, $\delta_2 < 1$ was used where $\theta_2 = \delta_2\phi_2$. Also, while imputing the continuous variables, it was assumed that those individuals lacking these variables are likely to be different from those predicted under the ignorable model by a factor of δ_3. Specifically, again dropping subscripts for brevity, suppose that Y are to be drawn from an 11-variable normal distribution with mean μ and covariance matrix Σ under the ignorable model. The imputations were created by drawing values from a normal distribution with mean $\delta_3\mu$ and the covariance matrix Σ.

TABLE 3

Adjusted odds ratios (and their 95% confidence limits) for five sets of values of the parameters for non-ignorability†

Parameters of non-ignorability	*Odds ratios for the following comparisons:*		
	Risk of thiazide compared with combined therapy (single drug)	*Risk of beta-blocker compared with combined therapy (single drug)*	*Risk of thiazide compared with combined therapy (multiple drug)*
$\delta_1 = 0.75, \delta_2 = 0.75, \delta_3 = 1.25$	1.08	0.75	0.78
	3.37	1.96	2.74
	8.11	8.36	8.69
$\delta_1 = 0.75, \delta_2 = 0.75, \delta_3 = 0.75$	1.50	1.11	1.02
	3.62	2.85	2.62
	8.12	9.94	8.25
$\delta_1 = 0.50, \delta_2 = 0.50, \delta_3 = 1.25$	1.70	1.12	1.01
	3.85	2.96	2.57
	7.02	8.31	7.98
$\delta_1 = 0.50, \delta_2 = 0.5, \delta_3 = 0.75$	1.52	1.16	1.02
	3.72	3.11	2.91
	7.08	8.03	8.26
$\delta_1 = 0.50, \delta_2 = 0.5, \delta_3 = 1.5$	1.64	1.22	1.12
	3.13	3.19	3.29
	7.29	8.74	8.79

†Number of cases, 164; number of controls, 742; the top entry in each group is the lower 95% confidence limit, the middle entry is the adjusted odds ratio and the bottom entry is the upper 95% confidence limit.

The results similar in format to Table 2 are summarized in Table 3 as a function of δ_1, δ_2 and δ_3. 15 imputations were used in all these cases. Apparently, there is some sensitivity to the ignorability assumption. The evidence in favour of combined therapy seems to be stronger in all the cases. The conclusion regarding the combined therapy alone when compared with thiazide alone remains the same at least qualitatively.

Another important component of the analysis is to explore the sensitivity with respect to the normality assumption of the continuous variables Y. The entire analysis just described was repeated assuming normality on the log-scale and also on the square-root scale. For the CIIS, which can take negative values, a large positive constant was added before taking logarithms or square-roots. The inferences obtained were very similar. At least in the example considered, the multiple-imputation inference seems to be less sensitive to normality than to the assumption of the ignorability of the missing data mechanism. This issue can be addressed more meaningfully by applying the method to simulated data sets generated under various distributional assumptions that deviate from the assumed normality.

7. Simulation Study

This section describes the results from a simulation study evaluating the sampling properties of the hybrid approach presented in this paper. The two objectives of the simulation study are

(a) to compare the sampling properties of the point and the interval estimates of the regression coefficient by using the hybrid multiple-imputation approach with those obtained by using the complete case and maximum likelihood approaches, and
(b) to investigate the robustness of the multiple-imputation approach to the distributional assumptions in the imputation model.

7.1. *Simulation Condition*

We considered a situation with binary disease (D) and exposure (E) variables that are known for all the individuals, a continuous confounding variable (Y) that may have missing values and a continuous variable (C) which is related only to Y and is also known for all the individuals. The ultimate user or the analyst is interested in fitting a logistic regression model with D as the dependent variable and E and Y as independent variables. The estimand of interest is the regression coefficient for E in the above logistic regression model.

A data set with 500 individuals was generated as follows. For each individual,

(a) C was generated as an independent standard normal deviate,
(b) Y was defined as $(1+\rho^2)^{-1/2}(C+\rho z)$ where z is an independent random variable,
(c) E was defined as an independent Bernoulli random variable with

$$\text{logit}(\Pr[E=1|C]) = -\log 4 + Y\log 3$$

and finally
(d) D was defined as an independent Bernoulli random variable with

$$\text{logit}(\Pr[D=1|E,\ Y]) = -\log 4 + E\log 3 + Y\log 2.$$

Thus the true value of the estimand is log 3. From these data, a data set with missing values was generated by deleting the value of Y based on a Bernoulli experiment with

$$\text{logit}(\Pr[\text{missing } Y|D,\ E]) = -\log 2 + E\log(2/7) + D\log 4.$$

This logit model for the missing data resulted in the approximate missing data percentages among exposed cases, exposed controls, unexposed cases and unexposed controls of 5.5%, 1.5%, 9% and 19% respectively.

Four possible distributions for the random variable z in step (b) above were considered:

(i) a normal distribution with mean 0 and variance 1,
(ii) a t-distribution with 3 degrees of freedom,
(iii) a translated log-normal distribution with mean 0 and variance 1 and
(iv) a translated exponential with mean 0 and variance 1.

Four possible values for $R^2 = 1/(1+\rho^2)$ of 0.8, 0.5, 0.3 and 0 were used to represent a range of predictive power of C for Y. For $R^2 = 0$, the random variables Y and C were generated as independent random variables. Thus this simulation study can be considered as a 4^2 factorial experiment. For each of the $4^2 = 16$ combinations, 10000 data sets with missing values were generated.

7.2. *Analysis*

Each data set was analysed by using three methods:

(a) the complete case analysis based on the individuals on whom Y is observed,
(b) the method of maximum likelihood as discussed by Little and Schluchter (1985) using a general location–scale model for the joint distribution of (D, E, Y) and
(c) the multiple-imputation analysis described in this paper.

The imputation model assumed that the conditional distribution of Y given C, $D=i$ and $E=j$ as a normal distribution with mean $\mu_{0ij}+\mu_{1ij}C$ and variance σ^2 where $\mu_{ij}=(\mu_{0ij}, \mu_{1ij})$ are independent bivariate normal random variables with mean (μ_0, μ_1) and covariance matrix Ω. As in the example, a diffuse prior distribution

$$\Pr(\mu_0, \mu_1, \sigma, \Omega) \propto \sigma^{-1}|\Omega|^{-2.25} \exp\{0.0001 \operatorname{tr}(\Omega^{-1})\}$$

was used. 50 imputations were created by drawing values from the posterior predictive distribution of Y given D, E and C via Gibbs sampling. Every 100th draw in 10 parallel Gibbs cycles was stored as an approximately independent draw. In each Gibbs cycle, the first 1000 draws were ignored to eliminate the effect of initial values. We chose $P=100$, although the Gelman–Rubin convergence statistic $\hat{R}$ (Gelman and Rubin (1992), p. 461) suggested that every 60th draw should be adequate. Similar results were obtained using $P=50$, 60, 80 and 90. Each completed data set was then analysed by fitting the logistic regression model relating D to E and Y. The completed data set estimates and the covariance matrices were combined as discussed in Section 5.

7.3. *Results*

The bias, the mean-square error and the exact coverage of the nominal 95% confidence interval for the three methods are displayed in Table 4. The mean-square errors of the three estimates are expressed as percentages of the mean-square errors of the estimates based on the complete or full data sets, i.e. before certain values were deleted to simulate data sets with missing values. Expressing the mean-square error as a percentage provides a useful yardstick to measure the loss of precision due to missing values for the three methods. Also, a comparison of these percentages for the multiple-imputation and the maximum likelihood methods allows us to measure the extent to which the loss of precision is mitigated by using the auxiliary variables.

Overall, the bias, the mean-square error and the confidence coverage are similar across the three distributions for z: normal, t_3 and log-normal. However, both the bias and the mean-square error are large and the exact coverage is smaller than the nominal coverage when the actual distribution of z is exponential. Thus both the multiple imputation and the maximum likelihood methods may be sensitive to extreme skewness of the actual distribution of the continuous variable.

The complete case estimates are severely biased and the exact coverages of the confidence intervals are far below the nominal level. This is expected because the data are not missing completely at random. Also, the complete case estimates have very large mean-square error because almost a third of the data have been ignored. Both the hybrid multiple imputation and the maximum likelihood estimates (ignoring C)

TABLE 4

Characteristics of three different estimates of the logistic regression coefficients for the simulated data sets†

Distribution of z	R^2*(%)*	*Characteristics*	*Results for the following methods:*		
			Complete case	*Maximum likelihood*	*Multiple imputation*
Normal	80	Bias	0.363	0.0091	0.0088
		MSE	460	145	108
		Coverage 95%	80	94	96
	50	Bias	0.401	0.0112	0.0089
		MSE	449	139	118
		Coverage 95%	78	95	96
	30	Bias	0.383	0.0082	0.0087
		MSE	450	142	127
		Coverage 95%	80	94	96
	0	Bias	0.393	0.0092	0.0089
		MSE	442	137	141
		Coverage 95%	80	95	96
t (3 degrees of freedom)	80	Bias	0.343	0.0082	0.0081
		MSE	421	148	107
		Coverage 95%	82	96	96
	50	Bias	0.386	0.0092	0.0095
		MSE	439	142	117
		Coverage 95%	81	96	96
	30	Bias	0.363	0.0072	0.0068
		MSE	442	139	128
		Coverage 95%	82	96	97
	0	Bias	0.383	0.0079	0.0073
		MSE	434	136	140
		Coverage 95%	80	95	97
Log-normal	80	Bias	0.399	0.0162	0.0172
		MSE	390	165	115
		Coverage 95%	80	94	94
	50	Bias	0.411	0.0152	0.0152
		MSE	449	169	126
		Coverage 95%	78	94	93
	30	Bias	0.393	0.0142	0.0129
		MSE	450	164	146
		Coverage 95%	80	94	94
	0	Bias	0.397	0.0152	0.0133
		MSE	448	166	171
		Coverage 95%	80	94	94
Exponential	80	Bias	0.443	0.0392	0.0399
		MSE	489	175	169
		Coverage 95%	80	92	93
	50	Bias	0.401	0.0302	0.0394
		MSE	449	185	176
		Coverage 95%	78	91	92
	30	Bias	0.399	0.0292	0.0399
		MSE	450	184	178
		Coverage 95%	80	92	92
	0	Bias	0.413	0.0402	0.0444
		MSE	462	199	203
		Coverage 95%	80	91	91

†MSE, mean-square error as a percentage of the mean-square error of the estimate for the complete data; coverage 95%, exact coverage of the nominal 95% confidence interval.

are almost unbiased. The hybrid multiple-imputation estimates are more efficient than the maximum likelihood estimates except when $R^2 = 0$. This is expected because when $R^2 = 0$ we are introducing noise in the imputed values and consequently the estimates are more variable. The nominal and the exact coverage are similar for both the multiple-imputation and the maximum likelihood interval estimates, although the multiple-imputation intervals are slightly conservative.

8. Discussion

We have developed a hybrid multiple-imputation approach where a Bayesian model based on an extended set of variables is used to create imputations, although the ultimate analysis may be based on a different set of variables. This approach is particularly useful if the same data were to be used by several investigators with differing statistical skills. Once the imputations have been carefully orchestrated, only some complete data analysis software and a module to combine the completed data inferences are needed. Of course, the adoption of the multiple-imputation technique is not paramount for borrowing strength from an extended set of variables. In some instances, the EM algorithm, for example, can be used to estimate the parameters of the larger model and then we can construct the estimates of the parameters in the smaller model from the estimates of the parameters in the larger model. The level of statistical sophistication required of the ultimate user to adopt this approach, however, is considerable.

The multiple-imputation approach can also be easily modified to explore sensitivity to the assumption about the ignorability of the missing data mechanism by creating imputations under a variety of non-ignorable missing data mechanisms. The approach used in the data analysis, though *ad hoc*, is appealing from a practical point of view and can be generalized to a more complicated setting. By combining the completed data inferences, where the imputations have been created under a variety of non-ignorable missing data mechanisms, we can incorporate the uncertainty due to the process creating the missing data.

An important issue that needs further attention is the effect of various choices of V_m, X_1 and U in the imputation scheme on the results of the ultimate user. If some important variables have been left out from V_m, X_1 or U, then it will introduce bias (due to underfitting) in the imputed values and which in turn may bias the estimates of the coefficients in the logistic model. In contrast, if irrelevant predictors are included in the model, then it can result in a loss of efficiency. In choosing the model, we may want to sacrifice efficiency for bias by using a big model although it may contain irrelevant predictors. We have also assumed that the covariance matrix Σ of Y is the same across all the cells formed by the categorical variables. The bias and the loss of efficiency may also result when this assumption is violated.

The model discussed in Section 3 may need modifications in certain instances. For example, the log-linear model for the cell probabilities may have to be modified to accommodate structural 0s. For instance, the question of frequency of smoking is asked only to those who are current smokers and hence the cell probability corresponding to the frequency of smoking is structurally 0 for non-smokers. There is also asymmetry in the model structure for the continuous and categorical variables. We have assumed a hierarchical model structure only for the cell means μ_m. But the estimation of the cell probabilities can be improved by imposing a similar random effect structure on, say, the logit of the probabilities especially with sparse data. This, however, increases the computational complexity. The example considered in this paper has only two binary variables with missing values (i.e. four cells in the multinomial part of the model) and hence the random effect structure was not considered. Also, an investigation of the residual deviances from the fitted log-linear model did not warrant the need for a random effect structure or an overdispersion parameter to explain any excess variability. Further studies are needed to develop algorithms and to investigate the improvement in the estimation process when random effect structures are imposed on both the cell means μ_m and the cell probabilities π.

Evidently, there is an increase in computational complexity and storage space required. All the computations discussed in this paper were performed on a 486 personal computer using GAUSS programming language (Aptech Systems, 1992). Alternative Monte Carlo methods can be used for drawing values from the posterior predictive distribution. We chose Gibbs sampling because of its ease of implementation for the particular model considered in this paper. Computational time may perhaps be saved by adopting more efficient strategies for drawing values. Finally, though the emphasis in this paper was on the analysis of data from a particular case–control study, the methods are easily extended to the analysis of data from a cohort study with binary, polytomous, count or continuous outcomes.

References

Aptech Systems (1992) *The GAUSS System Version 3.1.* Maple Valley: Aptech Systems.

Dellaportas, P. and Smith, A. F. M. (1993) Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Appl. Statist.*, **42**, 443–459.

DuMouchel, W. H. and Waternaux, C. E. (1992) Comment on "Hierarchical models for combining information and for meta analysis". In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 338–341. Oxford: Oxford University Press.

Gelfand, A. E. and Smith, A. M. F. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, **7**, 457–472.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.

Kleinbaum, D. G., Morgernstern, H. and Kupper, L. L. (1981) Selection bias in epidemiological studies. *Am. J. Epidem.*, **113**, 452–463.

Li, K. H., Raghunathan, T. E. and Rubin, D. B. (1991) Large sample significance levels from multiply imputed data using moment based statistics and an F reference distribution. *J. Am. Statist. Ass.*, **86**, 1065–1073.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data.* New York: Wiley.

Little, R. J. A. and Schluchter, M. D. (1985) Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, **72**, 497–512.

Louis, T. A. (1982) Finding the observed information when using the EM-algorithm. *J. R. Statist. Soc.* B, **44**, 226–233.

Olkin, I. and Tate, R. F. (1961) Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statist.*, **32**, 448–465.

Raftery, A. E. and Banfield, J. D. (1991) Comment on "Bayesian image restoration with two applications in spatial statistics". *Ann. Inst. Statist. Math.*, **43**, 32–43.

Raftery, A. E. and Lewis, S. (1992) How many iterations in Gibbs sampler? In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 763–773. Oxford: Oxford University Press.

Ritter, C. and Tanner, M. A. (1992) Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. *J. Am. Statist. Ass.*, **87**, 861–868.

Robins, J. M., Zhao, L. P., Rotnitzky, A. and Lipsitz, S. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846–866.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

——(1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. and Schenker, N. (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Statist. Ass.*, **81**, 366–374.

Schafer, J. L. (1994) *Analysis of Incomplete Multivariate Data by Simulation*. London: Chapman and Hall.

Schafer, J. L., Khare, M. and Ezzati-Rice, T. M. (1993) Multiple imputation of missing data in NHANES III. In *Proc. Bureau of Census 1993 A. Res. Conf.*, pp. 459–487. Washington DC: US Department of Commerce.

Siscovick, D. S., Raghunathan, T. E., Psaty, B. M., Koepsell, T. D., Wicklund, K. G., Lin, X., Cobb, L., Rautaharju, P. M., Copass, M. K. and Wagner, E. H. (1994) Diuretic therapy for hypertension and primary cardiac arrest. *New Engl. J. Med.*, **330**, 1852–1857.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.

Vach, W. and Blettner, M. (1991) Biased estimation of the odds ratio in case-control studies due to the use of *ad hoc* methods of correcting for missing values of confounding variables. *Am. J. Epidem.*, **134**, 895–907.

Vach, M. and Schumacher, M. (1993) Logistic regression with incompletely observed categorical covariates: a comparison of three approaches. *Biometrika*, **80**, 353–362.