Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 49 (2005) 821–836
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Bayesian modeling of missing data in clinical research

Peter C. Austin[a, b, c, *], Michael D. Escobar[b]

[a] *Institute for Clinical Evaluative Sciences, Toronto, Ont., Canada*
[b] *Department of Public Health Sciences, University of Toronto, Canada*
[c] *Department of Health Policy, Management and Evaluation, University of Toronto, Canada*

Received 7 June 2004; accepted 8 June 2004
Available online 1 July 2004

Abstract

The issue of missing data frequently confronts researchers using data derived from patient medical records. We used Monte Carlo simulations to examine the performance of three Bayesian methods that imputed missing data by placing a simple prior distribution upon the variable that was subject to being missing. These methods compared with two methods that used a multivariate logistic regression model to impute missing data. As a final comparator, we also examined the performance of the conventional complete case analysis and a method that equated missing data on a risk factor with evidence of absence of the risk factor. It was shown that the bias and mean square error of each method depended upon the prevalence of the risk factor under consideration and the missing data mechanism. No method performed well in all situations. However, assuming that the risk factor had a Bernoulli distribution and placing a uniform prior distribution upon the parameter of this distribution resulted in lower relative bias than the other competing methods in the majority of settings. The performance of each method was then examined on a dataset of 5131 patients admitted to hospital with a heart attack.

Keywords: Missing data; Clinical research; Bayesian analysis

* Corresponding author. Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ont., Canada M4N 3M5. Tel.: +1-416-480-6131; fax: +1-416-480-6048.
E-mail address: peter.austin@ices.on.ca (P.C. Austin)

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.06.006

1. Introduction

The issue of missing data frequently confronts researchers using data derived from patient medical records. Data abstracted from patient charts are frequently used for fitting regression models to determine the association between patient characteristics and adverse outcomes, and to construct valid risk-adjusted measures of outcomes, allowing for comparisons of outcomes between physicians or institutions. However, standard regression methods can only use patients for whom complete data are present, necessitating that patients with missing data be dropped from the analysis.

Ideally, in abstracting data from patient charts, the presence or absence of specific patient risk factors for the outcome of interest would be documented explicitly. For instance, if the presence of diabetes is a risk factor for which one would like to adjust in a regression model, then the medical record should state explicitly that either the patient had diabetes or that the patient did not have diabetes. A drawback to retrospectively extracting data from patient medical records is that the medical personnel updating a patient's medical chart are usually unaware of the future uses that will be made of the data. Hence, personnel updating the medical record may fail to explicitly state that the patient does not have diabetes. One solution is to gather data prospectively for specific research projects. However, in many instances, this is neither feasible nor cost effective.

Multiple methods, of varying sophistication, have been developed for handling missing data. Two naïve approaches are either to exclude subjects with missing data or to equate silence with absence. The first approach is referred to as a complete case analysis. The second approach assumes that if the medical record does not explicitly state that the patient had a given condition then, by default, the patient is assumed to not have had the condition. Hence "absence of evidence" is taken to imply "evidence of absence". While this second approach can be used when the variables under consideration are dichotomous variables indicating the presence or absence of a risk factor, it is not applicable in situations in which the variable is continuous (e.g. blood pressure, heart rate, hemoglobin level).

Vach (1994) examined missing data in the setting of a logistic regression model with two categorical predictor variables, only the second of which was subject to missing data. In this setting, if the probability of the second predictor being missing was independent of the outcome, then a complete case analysis results in consistent estimation of the regression coefficients, with only an associated loss of efficiency. Furthermore, if the likelihood of missing data is only related to the outcome variable, but not to any other covariate, then complete case analysis results in consistent estimation of the odds ratio, and only the intercept is changed (Vach, 1994). If the dependent variable is mortality, and if variables were more likely to missing for those who died, then a complete-case analysis would result in a downwards bias in the model intercept, since the sickest patients are systematically removed from the analysis. In a more general regression setting, Robins et al. (1994) state that if the probability of missing data is dependent only on completely observed predictor variables, then a complete case analysis will result in asymptotically unbiased estimates of the regression coefficients. However, if the probability of missing data depends on both the outcome variable and those variables that are completely observed, then complete case estimation may be inconsistent.

Another commonly used approach is to add a variable indicating whether a given variable was observed or whether it was missing. However, this approach has been shown to introduce bias in the estimated regression coefficients (Jones, 1996), and its use is discouraged (Vach, 1994). Furthermore, if it tends to be same patients who are missing multiple covariates, then collinearity will be introduced amongst the indicator variables.

The majority of methods for dealing with missing data use some form of imputing values for those that are missing, thus allowing all subjects to be used in the analysis. The simplest of these is mean-value imputation (Rubin, 1987; Little, 1992), in which the mean value of a given variable is imputed to all missing cases. While this approach is conceptually simple, its primary drawback is that it reduces the sampling variability in the data, since all individuals missing a given variable are imputed to have the same value for that variable. A secondary drawback to this approach is that it ignores the multivariate structure of the data. Furthermore, if data tend to be missing for the sickest patients, then mean-value imputation has the unintended drawback of making the sickest patients seem healthier than they truly are by imputing the mean values of patients with complete data, who tend to be healthier. Other approaches that attempt to preserve the multivariate structure of the data are "hot-deck" methods of imputation (Rubin, 1987). Much work has been carried out on developing multiple imputation methods that impute multiple values for each missing value. These methods aim to both preserve the multivariate structure of the data and to avoid artificially reducing the sampling variability (Rubin, 1987). The literature contains several reviews of methods for imputing missing data in general (Rubin, 1987; Little, 1992; Rubin and Schenker, 1991), and in epidemiology in particular (Greenland and Finkle, 1995). Rubin provides general conditions under which one can ignore the process under which missing data arose (Rubin, 1976).

The purpose of the current study was to compare simple Bayesian methods of imputation that place prior distributions upon variables that have missing data with that of an imputation method that uses the multivariate structure of the data. As a comparator, we also examine the traditional complete case analysis and the method that treats absence of evidence as evidence of absence. The first component of the study used Monte Carlo simulations to examine the impact of choice of prior probability distribution of the performance of the regression models. The second component of the study employed the methods examined earlier to examine the association between cardiac risk factors and mortality following admission to hospital for a heart attack.

2. Monte Carlo simulations: methods

2.1. Data generation

In the current study, we examined two data-generating processes. The first data-generating process assumed that a binary outcome, Y, was related to a continuous variable X_1 and a binary variable X_2. The first variable, X_1, was assumed to be observed for all subjects. This is frequently the case for variables such as age and gender that are available from computerized records or from demographic registries. It was assumed that the second variable, X_2, was subject to missing data. We are interested in making inferences about the association between

X_2 and the outcome, after adjusting for X_1. The following model was assumed:

$$Y_i \sim \text{Bernoulli}(p_i), \quad \text{where} \tag{1}$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}. \tag{2}$$

In particular, we assumed that $\beta_0 = 1$, $\beta_1 = 0.05$, and that $\beta_2 = 0.25$. It was assumed that $X_1 \sim N(0, \sigma = 10)$, and that $X_2 \sim \text{Bernoulli}(P_{\text{prevalence}})$, with 0 indicating absence of a condition and 1 indicating the presence of a condition, and with σ denoting the standard deviation of the normal distribution. One thousand observations were then generated using the described data-generating process.

The second data-generating process assumed that a binary outcome, Y, was related to a continuous variable X_1 and two binary variables X_2 and X_3. As above, the first variable, X_1, was observed for all subjects. It was assumed that both the binary variables were subject to missing data, and that they were either both present or both missing. Thus X_2 was missing if and only if X_3 was missing. We are interested in making inferences about the association between X_2 and the outcome, after adjusting for X_1 and X_3. The following model was assumed:

$$Y_i \sim \text{Bernoulli}(p_i), \quad \text{where} \tag{3}$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}. \tag{4}$$

In particular, we assumed that $\beta_0 = 1$, $\beta_1 = 0.05$, $\beta_2 = 0.25$, and that $\beta_3 = 0.25$. It was assumed that $X_1 \sim N(0, \sigma = 10)$, $X_2 \sim \text{Bernoulli}(P_{\text{prevalence}})$, and that $X_3 \sim \text{Bernoulli}(P_{\text{prevalence}})$, with 0 indicating absence of a condition and 1 indicating the presence of a condition, and with σ denoting the standard deviation of the normal distribution. Thus, for sake of simplifying the Monte Carlo simulations, it was assumed that the two dichotomous variables both had the same prevalence in the population. One thousand observations were then generated using the described data-generating process.

2.2. Missing data mechanism

The literature on missing data describes three different mechanisms by which missing data can arise: missing completely at random (MCAR), missing at random (MAR), and informative missing (IM) (Rubin, 1987). MCAR arises when the presence of missing data is unrelated to any subject characteristics. Data are said to be missing at random when the probability that a variable is missing is related to observed, non-missing variables. Informative missing occurs when the probability that a variable is missing is either related to the true value of the missing variable or to other unmeasured variables. The majority of methods for imputing missing data require the assumption that missing data are either MCAR or MAR. In this study, we shall focus on situations in which the missing data are IM. The rationale for this reflects our belief that absence of a risk factor is less likely to be noted in a patient's medical record than the presence of a risk factor. If the probability of a variable being missing is related only to that variable itself, then a complete case analysis will result in consistent estimation, albeit with a reduction in precision (Vach, 1994). We thus examined a mechanism for missing data in which the probability that the binary risk factor was missing was related to all variables, including the binary risk factor itself.

In the first data-generating process, we assumed that the probability of missing data was related to X_1, X_2, and to Y. In this setting, we defined $\text{logit}(P_{\text{missing}}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 Y$. We fixed α_1 to equal 0.05, and α_3 to equal 0.10. For each subject, we determined the probability that X_2 was missing, and then determined whether X_2 was missing for this subject by randomly generating a missing data indicator from a Bernoulli distribution with parameter P_{missing}. Hence data were more likely to be missing for those subjects who died ($Y = 1$), and for those subjects with increasing X_1 values. We used a full factorial design, allowing the following factors to vary: $P_{\text{prevalence}}$, α_0, and α_2. We allowed these factors to take the following levels:

$P_{\text{prevalence}} = 0.10$, 0.25, and 0.50,
$\alpha_0 = -1$, -2, and -3.
$\alpha_2 = -0.10$, and -0.20 (observations were less likely to be missing if the binary risk factor was truly present).

For each of the 18 scenarios, we generated a dataset of 1000 subjects and fit seven different models to the data. This procedure was repeated 100 times for each scenario.

In the second data-generating process, we assumed that the probability of missing data was related to X_1, X_2, X_3, and to Y. In this setting, we defined $\text{logit}(P_{\text{missing}}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_2 X_3 + \alpha_3 Y$. We fixed α_1 to equal 0.05, and α_3 to equal 0.10. Hence data were more likely to be missing for those subjects who died ($Y = 1$), and for those subjects with increasing X_1 values. We used a full factorial design, allowing the following factors to vary: $P_{\text{prevalence}}$, α_0, and α_2. We allowed these factors to take the same levels as in the above design. As above, or each of the 18 scenarios, we generated a dataset of 1000 subjects and fit seven different models to the data. This procedure was repeated 100 times for each scenario.

3. Statistical models

Seven different regression models, each making different assumptions about missing data, were fit to the randomly generated data.

Model 1: $Y_i \sim \text{Bernoulli}(p_i)$,
$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,
Subjects with missing data are dropped prior to the analysis. Thus, our first method is the conventional complete case analysis.

Model 2: $Y_i \sim \text{Bernoulli}(p_i)$,
$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,
If X_2 is missing then it is coded as 0 prior to the analysis.
This model equates missing values with the absence of the condition. This method does not have an explicit model for the missing data, but simply equates silence about the condition with evidence of absence of the condition.

Model 3: $Y_i \sim \text{Bernoulli}(p_i)$,
$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,
$X_2 \sim \text{Bernoulli}(P_{\text{empirical},2})$,
where $P_{\text{empirical},2}$ denotes the observed proportion of patients for whom $X_2 = 1$, amongst those patients for whom X_2 is not missing. This model assumes that if X_2 is missing then the probability that its true value is 1 is equal to the observed proportion of patients with $X_2 = 1$.

Model 4: $Y_i \sim \text{Bernoulli}(p_i)$,
$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,
$X_2 \sim \text{Bernoulli}(P_2)$,
$P_2 \sim \text{Beta}(1, 1)$. This model assumes that X_2 follows a Bernoulli distribution, and a uniform prior is placed upon the parameter of the Bernoulli distribution. This is the first model to incorporate uncertainty into the true probability that $X_2 = 1$.

Model 5: $Y_i \sim \text{Bernoulli}(p_i)$,
$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,
$X_2 \sim \text{Bernoulli}(P_2)$,
$P_2 \sim \text{Beta}(\alpha_2, \beta_2)$, where α_2 and β_2 are such that $\alpha_2 + \beta_2 = 100$ and $\alpha_2/(\alpha_2 + \beta_2)$ equals the proportion of subjects with $X_2 = 1$ among those patients with no missing data. This model incorporates uncertainty into the true probability that $X_2 = 1$. The parameters α_2 and β_2 were chosen so that the mean of the prior Beta distribution was equal to the observed proportion of patients for whom $X_2 = 1$, amongst those patients for whom X_2 is not missing, and so that the weight of the prior belief was equivalent to observing 100 subjects.

Model 6: $Y_i \sim \text{Bernoulli}(p_i)$,
$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,
$X_2 \sim \text{Bernoulli}(P_i)$,
$\text{logit}(P_i) = \alpha_0 + \alpha_1 X_1 + \alpha_2 Y_2$, and
$\alpha_i \sim N(0, \sigma = 5)$.

Model 7: $Y_i \sim \text{Bernoulli}(p_i)$,
$\text{logit}(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,
$X_2 \sim \text{Bernoulli}(P_i)$,
$\text{logit}(P_i) = \alpha_0 + \alpha_1 X_1 + \alpha_2 Y_2 + \alpha_3 X_1 Y_2$ and
$\alpha_i \sim N(0, \sigma = 5)$.
This model is a modification of Model 6 in which the logistic regression model for imputation is mis-specificied.

For each of the six models, proper priors were used for the regression parameters: $\beta_0 \sim N(1, \sigma = 1)$, $\beta_1 \sim N(0.05, \sigma = 0.025)$, and $\beta_2 \sim N(0.25, \sigma = 0.125)$. The means of the

Table 1
Burn-in and monitored iterations for the MCMC process

Model	First data-generation process		Second data-generation process	
	Burn-in iterations	Monitored iterations	Burn-in iterations	Monitored iterations
Model 1	1500	500	1500	500
Model 2	1000	500	1000	500
Model 3	500	500	500	500
Model 4	1500	500	1500	500
Model 5	500	500	500	500
Model 6	1600	1000	1600	500
Model 7	2500	500	3000	500

prior distributions were chosen to be equal to the parameters used in the data-generation process in order to speed the convergence of the Markov chain Monte Carlo algorithm.

The models described above are for the first data-generating process, in which the log-odds of the outcome were linearly related to X_1 and X_2. In the second data-generating process, the log-odds of the outcome were linearly related to X_1, X_2 and X_3. Each of the above seven models was easily adapted to the second data-generating process by adding a term for X_3 in each of the regression models (with associated regression parameter β_3). Similarly, the distribution assumed for X_3 would be the same as that assumed for X_2. Finally, the same prior distribution would be assumed for β_3 as for β_2.

Each of the above seven models was fit using the Markov chain Monte Carlo methods (Gilks et al., 1996) using BUGS (Bayesian Inference Using Gibbs Sampling) (Gilks et al., 1994). Each model was fit to the randomly generated data. Since 100 random datasets were generated for each of 18 scenarios and for each of two data-generating processes, exploratory analyses were performed to determine the required number of "burn-in" iterations for the Gibbs sampler for each of the seven models. For each of the two data-generating processes (one dichotomous predictor versus two dichotomous predictors), a random dataset was generated. The Gibbs sampler was run to estimate each of the seven models and the chain was then monitored. The monitored chains were then inspected visually to assess whether the Gibbs sampler had attained stationarity. In addition, Geweke's statistic (Geweke, 1994) was used to assess whether the monitored chain had attained stationarity. Once an appropriate number of burn-in iterations had been determined for a particular model and data-generating process, this number of burn-in iterations was used for all estimation of that model. The number of burn-in iterations and length of the monitored chain for each model and for each data-generating process are described in Table 1 .

The mean regression parameters from the sampled chains were computed for each of the regression parameters. This process was repeated 100 times for each of the 18 scenarios in each of the two data-generating processes. For each scenario, the mean square error (MSE) and bias in estimating β_2 was computed.

Table 2
Relative bias in estimated regression coefficient for X_2 (first data-generating process)

Prevalence of risk factor	$a0$	$a2$	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
0.1	−3	−0.2	1.8	−1.4	1.8	1.1	2.1	2.3	2.2
0.1	−3	−0.1	2.2	−2.4	0.9	0.5	1.3	1.9	1.7
0.1	−2	−0.2	2.7	−5	1.4	1.5	1.7	4.4	4.9
0.1	−2	−0.1	1.3	−2.6	0.5	0.8	0.5	3.3	3.4
0.1	−1	−0.2	2.3	−1.4	−0.4	0	0.1	6.2	6.4
0.1	−1	−0.1	−0.2	−2.2	0.4	0.1	0	7	7
0.25	−3	−0.2	4.2	−1.7	5.7	5.1	5.9	7.4	7.7
0.25	−3	−0.1	−3.1	1.3	4.6	4.2	4.9	6.4	6.4
0.25	−2	−0.2	−1.9	−3.9	1.7	1.4	1.8	6.1	6.6
0.25	−2	−0.1	−3.2	−2.4	2.4	2.4	2.1	7.2	7.3
0.25	−1	−0.2	1.5	−7.2	−0.5	−0.1	−0.4	11.7	12.4
0.25	−1	−0.1	−0.5	−4.7	−7.6	−7.9	−7.7	1.6	1.9
0.5	−3	−0.2	0	1	−0.5	−1.1	0.1	1.7	1.3
0.5	−3	−0.1	0.2	−1.5	−1.5	−1.7	−0.9	0.7	0.5
0.5	−2	−0.2	0.9	−4.7	3.2	3.2	3.2	8.6	9.3
0.5	−2	−0.1	−1.7	−4	4.5	4.8	4.8	10.9	11.3
0.5	−1	−0.2	−1.6	−15	−1.2	−1.1	−1.6	12.6	12.7
0.5	−1	−0.1	4.5	−11.6	−2.8	−2.7	−2.8	11.9	12.4

Each cell contains the estimated relative bias in the given combination of the factors.

4. Monte Carlo simulations: results

4.1. First data-generating process (one binary variable subject to missing data)

The estimated relative bias and MSE in each of the 18 different scenarios are reported in Tables 2 and 3 respectively. Models 6 and 7 had the largest relative biases in 15 of the 18 scenarios examined. Models 3–5 tended to have the lowest absolute relative biases. However, there were scenarios in which model 1 (complete case analysis) or model 2 (naïve model assuming that missing implies absence) had the lowest relative bias. In examining Table 3, one observes that none of the models had the lowest MSE in all 18 scenarios. However, models 6 and 7 had the highest MSE in 13 of the 18 scenarios examined. In contrast to this, models 3–5 had the lowest MSE in a majority of scenarios.

4.2. Second data-generating process (two binary variables subject to missing data)

The estimated relative bias and MSE in each of the 18 different scenarios are reported in Tables 4 and 5 respectively. No model consistently outperformed the other models. Models 2, 6, and 7 tended to have the greatest relative bias, while models 3–5 tended to have the lowest relative bias. Similarly, no model consistently had the lowest MSE. However, models 6 and 7 tended to have the highest MSE.

Table 3
MSE in estimated regression coefficient for X_2 (first data-generating process)

Prevalence of risk factor	$a0$	$a2$	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
0.1	−3	−0.2	0.003	0.002	0.002	0.002	0.002	0.002	0.002
0.1	−3	−0.1	0.002	0.003	0.002	0.002	0.002	0.002	0.002
0.1	−2	−0.2	0.002	0.003	0.002	0.002	0.002	0.002	0.002
0.1	−2	−0.1	0.003	0.002	0.003	0.003	0.003	0.003	0.003
0.1	−1	−0.2	0.002	0.002	0.002	0.002	0.002	0.004	0.004
0.1	−1	−0.1	0.002	0.002	0.002	0.002	0.002	0.004	0.004
0.25	−3	−0.2	0.004	0.003	0.004	0.004	0.004	0.005	0.005
0.25	−3	−0.1	0.004	0.004	0.004	0.004	0.004	0.004	0.004
0.25	−2	−0.2	0.004	0.003	0.003	0.003	0.003	0.004	0.004
0.25	−2	−0.1	0.003	0.005	0.003	0.003	0.003	0.004	0.004
0.25	−1	−0.2	0.003	0.004	0.002	0.002	0.002	0.005	0.005
0.25	−1	−0.1	0.003	0.004	0.003	0.004	0.003	0.006	0.006
0.5	−3	−0.2	0.003	0.004	0.005	0.005	0.005	0.006	0.006
0.5	−3	−0.1	0.005	0.003	0.003	0.003	0.003	0.004	0.004
0.5	−2	−0.2	0.004	0.003	0.005	0.005	0.005	0.006	0.007
0.5	−2	−0.1	0.003	0.004	0.003	0.003	0.003	0.005	0.005
0.5	−1	−0.2	0.004	0.007	0.004	0.004	0.004	0.009	0.009
0.5	−1	−0.1	0.004	0.005	0.004	0.004	0.004	0.008	0.009

Each cell contains the estimated MSE in the given combination of the factors.

5. Case study

5.1. Data sources

Data for the case study was obtained by examining the medical records of 5131 patients who were admitted to 44 hospitals in Ontario, Canada with a diagnosis of a heart attack or acute myocardial infarction (AMI). Data on demographic characteristics, cardiac risk factors and the clinical profile of each patient was abstracted by examining the patients' medical records. Data on each patient's vital status was obtained by linking data from the medical record to the Registered Persons Database (RPDB), which records the vital status of all Ontario residents. Data on vital status, age and gender were present for all patients.

5.2. Methods

The association between history of hyperlipidemia, smoking status, and family history of heart disease and mortality within 30 days of admission to hospital was determined. Data on hyperlipidemia, smoking history, and family history of heart disease were missing for 2.5%, 9.8%, and 14.6% of the patients, respectively. Each of the models described in the Monte Carlo simulations was fit to the data to determine the association between each cardiac risk factor and 30-day mortality, after adjusting for age and gender. The logistic

Table 4
Relative bias in estimated regression coefficient for X_2 (second data-generating process)

Prevalence of risk factor	$a0$	$a2$	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
0.1	−3	−0.2	0.8	4.1	0.5	0.8	0.4	1.5	1.7
0.1	−3	−0.1	0.2	1.5	0.2	0.2	0.4	1.2	1.3
0.1	−2	−0.2	4.3	1.1	−1.2	−1.0	−1.7	1.0	1.1
0.1	−2	−0.1	3.1	−4.7	3.3	3.2	3.0	6.5	5.8
0.1	−1	−0.2	0.0	−1.5	2.5	1.8	1.8	8.8	8.9
0.1	−1	−0.1	3.4	−5.8	−1.6	−2.1	−1.7	5.1	5.5
0.25	−3	−0.2	4.5	0.9	−4.3	−3.8	−4.1	−2.8	−2.4
0.25	−3	−0.1	5.3	−1.4	−1.0	−0.7	−0.9	0.3	0.4
0.25	−2	−0.2	4.3	−3.2	4.5	4.1	4.6	8.6	8.8
0.25	−2	−0.1	1.8	−2.4	1.4	1.5	1.5	5.9	5.9
0.25	−1	−0.2	0.1	−7.7	0.0	0.3	−0.4	11.5	12.0
0.25	−1	−0.1	1.9	−4.7	0.0	−0.3	−0.6	10.2	11.0
0.5	−3	−0.2	3.5	−8.8	1.1	1.1	1.0	2.6	3.3
0.5	−3	−0.1	1.9	−2.9	−0.8	−0.6	−0.8	1.2	1.1
0.5	−2	−0.2	0.8	−7.8	0.0	−0.4	−0.2	4.5	4.3
0.5	−2	−0.1	2.2	−7.8	4.0	3.4	4.0	8.6	9.0
0.5	−1	−0.2	3.2	−12.8	0.3	0.4	1.0	14.5	14.5
0.5	−1	−0.1	0.1	−16.8	−0.4	−0.6	−0.4	14.7	15.3

Each cell contains the estimated relative bias in the given combination of the factors.

regression model that was fit was:

$$\text{logit}(p_i) = \beta_0 + \beta_1\ \text{Age}_i + \beta_2\ \text{Gender}_i + \beta_3 X_{3i}, \tag{5}$$

where X_3 denotes a risk factor (either hyperlipidemia, smoking history, or family history of heart disease) for the ith patient. Thus, three separate regression models were fit, each of which determined the association between a particular risk factor and mortality, after adjusting for only age and gender. Furthermore, it should be noted that the association between a given risk factor and mortality was determined after adjustment for age and gender, but not for the other risk factors.

The 30-day mortality rate stratified by each level of each risk is reported in Table 6. Females had a substantially higher risk of death within 30-days than did males. Current smokers, those with a history of hyperlipidemia, and those with a family history of heart disease had a lower rate of 30-day mortality than those who did not. However, for each of the three risk factors, the highest mortality rate was observed for those who had missing data for the given risk factor. Table 7 summarizes the age distribution at each level of each risk factor. Documented current smokers were substantially younger than patients who were documented non-smokers. However, patients with missing data on current smoking status were older than either group. Patients with a history of hyperlipidemia were younger than those with a documented absence of hyperlipidemia. However, patients with missing data were older than both of these groups. Similarly, patients with a documented family history of heart disease were younger than patients with a documented absence of family history

Table 5
MSE in estimated regression coefficient for X_2 (second data-generating process)

Prevalence of risk factor	$a0$	$a2$	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
0.1	−3	−0.2	0.002	0.002	0.002	0.002	0.002	0.002	0.002
0.1	−3	−0.1	0.003	0.002	0.002	0.002	0.002	0.002	0.002
0.1	−2	−0.2	0.003	0.002	0.002	0.002	0.002	0.003	0.003
0.1	−2	−0.1	0.003	0.002	0.003	0.003	0.003	0.003	0.003
0.1	−1	−0.2	0.002	0.002	0.002	0.002	0.002	0.003	0.003
0.1	−1	−0.1	0.002	0.002	0.002	0.002	0.002	0.003	0.003
0.25	−3	−0.2	0.003	0.004	0.004	0.004	0.004	0.004	0.004
0.25	−3	−0.1	0.004	0.003	0.003	0.003	0.003	0.003	0.003
0.25	−2	−0.2	0.003	0.004	0.005	0.005	0.005	0.006	0.006
0.25	−2	−0.1	0.003	0.004	0.004	0.004	0.004	0.005	0.005
0.25	−1	−0.2	0.003	0.004	0.002	0.003	0.002	0.004	0.004
0.25	−1	−0.1	0.003	0.003	0.004	0.004	0.004	0.006	0.006
0.5	−3	−0.2	0.004	0.005	0.004	0.004	0.004	0.004	0.004
0.5	−3	−0.1	0.004	0.004	0.004	0.004	0.004	0.004	0.004
0.5	−2	−0.2	0.003	0.004	0.004	0.004	0.004	0.004	0.004
0.5	−2	−0.1	0.003	0.004	0.004	0.004	0.004	0.005	0.005
0.5	−1	−0.2	0.004	0.005	0.004	0.004	0.004	0.006	0.006
0.5	−1	−0.1	0.003	0.005	0.003	0.003	0.003	0.005	0.006

Each cell contains the estimated MSE in the given combination of the factors.

Table 6
30-day mortality rates for each stratum of each risk factor

Risk factor	Yes	No	Missing
Female gender	17.4%	11.5%	NA
Current smoker	9.0%	16.4%	27.0%
Hyperlipidemia	6.0%	15.1%	27.0%
Family history of heart disease	5.6%	15.3%	25.3%

of heart disease. Again, patients with missing data on family history of heart disease were older than either group.

The models were estimated using Markov chain Monte Carlo methods using the BUGS software program. Proper priors were assumed for each of the regression parameters.

5.3. Results

The odds ratio for the association between hyperlipidemia, smoking history, and family history of heart disease and 30-day mortality, along with 95% credible intervals are summarized in Table 8.

Table 7
Age distribution within each stratum of each risk factor

Risk factor (level)	Mean	1st quartile	Median	3rd quartile
Smoker:				
No	71.8	65	73	81
Yes	62.8	53	64	72
Missing	73.8	68	75	82
Hyperlipidemia:				
No	68.6	60	71	78
Yes	61.9	53	63	71
Missing	71.6	65	72	79
Family history:				
No	69.3	62	71	78
Yes	61.6	52	62	71
Missing	73.4	67	74	82

Table 8
Association between cardiac risk factors and 30-day mortality

Risk factor	Hyperlipidemia	Smoking history	Family history of heart disease
Model 1	0.54 (0.41, 0.72)	0.83 (0.68, 0.99)	0.48 (0.38, 0.60)
Model 2	0.53 (0.39, 0.69)	0.72 (0.60, 0.87)	0.43 (0.34, 0.54)
Model 3	0.55 (0.41, 0.71)	0.87 (0.71, 0.93)	0.52 (0.41, 0.66)
Model 4	0.55 (0.42, 0.72)	0.88 (0.72, 0.94)	0.53 (0.42, 0.66)
Model 5	0.55 (0.40, 0.71)	0.88 (0.72, 0.94)	0.52 (0.41, 0.64)
Model 6	0.53 (0.40, 0.69)	0.78 (0.64, 0.94)	0.39 (0.31, 0.49)
Model 7	0.53 (0.40, 0.69)	0.79 (0.65, 0.95)	0.39 (0.30, 0.50)

Each cell contains the odds ratio (95% credible interval) for the association between the risk factor in question and 30-day mortality (after adjusting for age and gender).

Previous history of hyperlipidemia had the least missing data (2.5% of subjects). The seven models for determining the age and sex-adjusted association between 30-day mortality and history of hyperlipidemia all produced comparable results. Using all seven methods, hyperlipidemia was associated with decreased mortality. The odds ratios and associated 95% credible intervals were similar across models.

Smoking history was missing for 9.8% of the patients. All seven models found that, after adjusting for age and gender, smoking was associated with decreased 30-day mortality. However, the strength of the association differed across models. The method that treated missing data on smoking status as evidence of no smoking resulted in the strongest association between smoking and decreased mortality. The three models that placed simple priors upon the distribution of the binary risk factor all produced similar results, and also produced the weakest association between smoking and short-term mortality. The two methods that used multivariate models for imputing smoking status resulted in similar conclusions.

Our finding that smoking was associated with decreased mortality was surprising, but has been observed in earlier studies. In analyzing data from the Thrombolysis in Myocardial Infarction (TIMI) Trial, Phase II, Mueller et al. (1992) observed that current smokers and ex-smokers had a lower mortality rate than never-smokers. This finding disappeared upon multivariate adjustment using regression models. The lower mortality of smokers seemed to be due to the lower risk profile of current smokers. Never-smokers tended to be older, were more likely to be female, and more frequently had pulmonary edema or cardiogenic shock, and were more likely to have had a history of diabetes and congestive heart failure (Mueller, 1992). Thus, it is likely that residual confounding existed in our analyses, and that the observed association between smoking and mortality would disappear upon further adjustment.

Family history of heart disease was missing for 14.5% of the patients. All seven models found that, after adjusting for age and gender, family history of heart disease was associated with decreased 30-day mortality. However, the strength of the association differed across models. The three models that placed simple priors upon the distribution of the binary risk factor all produced similar results, and also produced the weakest association between family history and short-term mortality. The two methods that used a multiple regression model to impute missing data produced similar results and also resulted in the largest effect size.

6. Discussion

We compared the performance of seven methods for dealing with missing data. The first method was the commonly used complete case analysis; the second equated missing data with the absence of the risk factor; three methods involved placing a simple prior distribution on the parameter with missing data; and two methods used multiple logistic regression models to impute values for missing data. Each model was assessed in two different settings. In the first setting, the probability that the binary risk factor was missing was related to the binary risk factor, a continuous covariate and the dichotomous outcome. In the second setting, an additional dichotomous predictor was included that was subject to missing data as well. In both settings data were subject to informative missingness (IM). Hence, traditional imputation methods were inappropriate.

Model performance was assessed by computing the bias and mean square error (MSE) in estimating the logarithm of the odds ratio associated with the binary risk factor. The seven methods that we examined can be grouped into three broad categories. The first category consisted of two methods. The first method was the conventional complete case analysis, while the second method equated missing data with absence of the risk factor. Neither of these methods had an explicit model for the missing data. The second category consisted of three models. Each of these models placed a Bernoulli prior distribution on distribution of the risk factor. The first model assumed that the prevalence of the risk factor was equal to the observed prevalence of the risk factor amongst those patients who did not have missing data. The second and third models in this category placed Beta hyperpriors upon the parameter of the Bernoulli distribution. The third category consisted of two models, each of which used a multivariate model to impute missing values. In general, models in a given category

had similar performance to the other models in the same category. There was no model that was clearly favored in all situations. However, the models in the second category tended to have superior performance, although there were several scenarios in which the seemingly naïve models in the first category had relatively better performance than the other models. The two models that used a multiple logistic regression model to impute missing values, and hence preserve the multivariate structure of the data, resulted in both higher MSE and bias than the other five models in most scenarios examined. We hypothesize that this was due to the fact that the imputation models used were mis-specified. The probability that the binary risk factor was missing was related to all variables, including itself. However, the imputation model was not able to include the binary risk factor as a predictor. We examined two settings: one in which only one dichotomous variable was subject to missing data and a second in which two dichotomous variables were both subject to missing data. Models in the second category of methods tended to perform well in both data-generating processes. Similarly, the two methods in the third category of models tended to perform poorly in both data-generating processes. However, the naïve method that equated missingness with absence tended to have poor performance relative to the competing methods in the setting with two dichotomous variables, both of which were subject to missing data. This was not the case in the first data-generating process.

None of the methods examined performed perfectly in all settings. However, we can make two recommendations. First, in the settings that we examined, the use of the multivariate imputation methods should be avoided. When the likelihood that the binary risk factor is missing is related to all variables, including the risk factor itself, the imputation model will be mis-specified. Second, the models that placed a Bernoulli prior distribution on prevalence of the risk factor tended to have lower relative bias compared to the other methods. While these methods did not result in unbiased estimation, they tended to have superior performance than the other models. Thus, we would recommend their use over the other methods examined in this study.

Using data on 5131 patients hospitalized with an acute myocardial infarction (AMI), we used the seven models described above to determine the association between cardiac risk factors and 30-day mortality after adjusting for age and gender. When the prevalence of missing data was low, the seven models produced comparable results. However, as the prevalence of missing data increased, divergent results were obtained. Equating missing risk factors with evidence that the risk factor was absent for the patient in question tended to magnify the strength of the association between the risk factor and mortality, compared to those models that placed simple priors upon the prevalence of the risk factor in the population.

Bayesian methods for analyzing data subject to missing values have been used previously in the literature, particularly for the analysis of longitudinal clinical trials with loss to follow-up. Cowles et al. (1996) implemented a Bayesian Tobit model for longitudinal ordinal clinical trial data with nonignorable missingness. Carpenter et al. (2002) used Bayesian methods to model both the response to treatment and the drop-out process in an asthma clinical trial. West and Dawson (2002) used Bayesian methods to analyze data comprised of incomplete categorical data in longitudinal studies. Ramachandran and Vincent (1999), using limited historical measurements and Bayesian methods, obtained estimates of exposure histories for airborne particulates, allowing for the estimation of a dose–response

relationship between exposure to pollutants and adverse health effects. Similarly, other authors have fit Bayesian models to account for loss to follow-up and discontinuation of treatment in clinical trials (Kleinman et al., 1998; Lavori et al., 1995). In an application, similar to the current, Kadane used a Bayesian analysis for surveys with missing data (Kadane, 1993). However, in his analysis, he examined not missing data on select variables, but subjects for whom all data were missing due to survey non-response. Chiu and Sedransk (1986) developed a Bayesian procedure for imputing missing values in sample surveys in which prior information from earlier surveys could be incorporated.

In this study, we have used Monte Carlo simulations to determine the bias and mean squared error associated with different Bayesian methods for imputing missing data in clinical research. Bayesian analysis rests upon determining the posterior distribution of the parameter of interest. Until the advent of Markov chain Monte Carlo methods, Bayesian analyses were restricted to simple models in which the likelihood could be explicitly determined. Due to the analytic difficulty of determining the posterior distribution, it is unlikely that the results presented in this study could have been determined analytically. Hence, Monte Carlo simulations were used to assess bias and mean squared error.

There has been limited research into the utility of different Bayesian models for imputing missing data in clinical research. The results of the current study suggest that placing simple prior distributions on the binary variables that are subject to missing data results in relatively better performance than simpler, more naïve approaches or than multivariate imputation in many settings.

Acknowledgements

The Institute for Clinical Evaluative Sciences (ICES) is supported in part by a grant from the Ontario Ministry of Health and Long Term Care. The opinions, results and conclusions are those of the author and no endorsement by the Ministry of Health and Long-Term Care or by the Institute for Clinical Evaluative Sciences is intended or should be inferred. Dr. Austin is supported in part by a New Investigator award from the Canadian Institutes of Health Research (Institute for Health Services and Policy Research). This research was supported in part by operating grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Carpenter, J., Pocock, S., Johan, L.C., 2002. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. Statist. Med. 21, 1043–1066.

Chiu, H.Y., Sedransk, J., 1986. A Bayesian procedure for imputing missing values in sample surveys. J. Amer. Statist. Assoc. 81, 667–676.

Cowles, M.K., Carlin, B.P., Connett, J.E., 1996. Bayesian Tobit modeling of longitudinal ordinal clinical trial data with nonignorable missingness. J. Amer. Statist. Assoc. 91, 86–98.

Geweke, J., 1994. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 4. Clarendon Press, Oxford, pp. 169–193.

Gilks, W.R., Thomas, A., Spiegelhalter, D.J., 1994. A language and program for complex Bayesian modelling. Statistician 43, 169–178.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Introducing Markov chain Monte Carlo. in: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London, pp. 1–19.

Greenland, S., Finkle, W.D., 1995. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Amer. J. Epidemiol. 142, 1255–1264.

Jones, M.P., 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. J. Amer. Statist. Assoc. 91, 222–230.

Kadane, J.B., 1993. Subjective Bayesian analysis for surveys with missing data. Statistician 42, 415–426.

Kleinman, K.P., Ibrahim, J.G., Laird, N.M., 1998. A Bayesian framework for intent-to-treat analysis with missing data. Biometrics 54, 265–278.

Lavori, P.W., Dawson, R., Shera, D., 1995. A multiple imputation strategy for clinical trials with truncation of patient data. Statist. Med. 14, 1913–1925.

Little, R.J.A., 1992. Regression with missing *X*'s: a review. J. Amer. Statist. Assoc. 87, 1227–1237.

Mueller, H.S., Cohen, L.S., Braunwald, E., Forman, S., Feit, F., Ross, A., Schweiger, M., Cabin, H., Davison, R., Miller, D., Solomon, R., Knatterud, G.L., 1992. Predictors of early morbidity and mortality after thrombolytic therapy of acute myocardial infarction: analyses of patient subgroups in the thrombolysis in myocardial infarction (TIMI) trial, phase II. Circulation 85, 1254–1264.

Ramachandran, G., Vincent, J.H., 1999. A Bayesian approach to retrospective exposure assessment. Appl. Occup. Environ. Hyg. 14, 547–557.

Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 89, 846–866.

Rubin, D.B., 1976. Inference and missing data. Biometrika 3, 581–592.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York, NY.

Rubin, D.B., Schenker, N., 1991. Multiple imputation in health-care databases: an overview and some applications. Statist. Med. 10, 585–598.

Vach, W., 1994. Logistic Regression with Missing Values in the Covariates. Springer, New York, NY.

West, C.P., Dawson, J.D., 2002. Complete imputation of missing repeated categorical data: one-sample applications. Statist. Med. 21, 203–217.