VOLUME 30 · NUMBER 26 · SEPTEMBER 10 2012

JOURNAL OF CLINICAL ONCOLOGY

SPECIAL ARTICLE

Missing Data in Clinical Studies: Issues and Methods

Joseph G. Ibrahim, Haitao Chu, and Ming-Hui Chen

Joseph G. Ibrahim, University of North Carolina at Chapel Hill, Chapel Hill, NC; Haitao Chu, University of Minnesota, Minneapolis, MN; and Ming-Hui Chen, University of Connecticut, Storrs, CT.

Submitted August 5, 2011; accepted March 14, 2012; published online ahead of print at www.jco.org on May 29, 2012.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Joseph G. Ibrahim, PhD, Department of Biostatistics, University of North Carolina, CB # 7420, Chapel Hill, NC 27599; e-mail: ibrahim@bios.unc.edu.

0732-183X/12/3026-3297/$20.00

DOI: 10.1200/JCO.2011.38.7589

ABSTRACT

Missing data are a prevailing problem in any type of data analyses. A participant variable is considered missing if the value of the variable (outcome or covariate) for the participant is not observed. In this article, various issues in analyzing studies with missing data are discussed. Particularly, we focus on missing response and/or covariate data for studies with discrete, continuous, or time-to-event end points in which generalized linear models, models for longitudinal data such as generalized linear mixed effects models, or Cox regression models are used. We discuss various classifications of missing data that may arise in a study and demonstrate in several situations that the commonly used method of throwing out all participants with any missing data may lead to incorrect results and conclusions. The methods described are applied to data from an Eastern Cooperative Oncology Group phase II clinical trial of liver cancer and a phase III clinical trial of advanced non–small-cell lung cancer. Although the main area of application discussed here is cancer, the issues and methods we discuss apply to any type of study.

J Clin Oncol 30:3297-3303. © 2012 by American Society of Clinical Oncology

INTRODUCTION

In clinical trials, common end points include response rate, immune response, quality of life (QOL), and survival time. Investigators often wish to quantify the effects of prognostic factors (covariates) on the outcome to adjust for imbalances in covariates that may persist after randomization or to model the natural history of the disease. Longitudinally modeling a patient self-reported outcome, such as QOL, is also often of interest. Unfortunately, missing outcome and/or covariate data are a common occurrence for most medical studies. Generally speaking, a participant variable can be regarded as missing if the value of the variable (outcome or covariate) for the participant is not observed. Missing data fractions may be large for some studies, especially for studies in which the covariate consists of a laboratory measurement or biomarker that is difficult to measure or for longitudinal studies in which there is heavy study dropout because of treatment toxicity.[1,2] The omission of participants with missing values can have a big impact on the analysis.[3–5] Measurements that require confidential, invasive, or painful collection procedures, complicated laboratory analysis, and/or time-consuming coding or compilation are more likely to be missing. The missing data can be in the form of missing outcome data, such as missing QOL measurements in longitudinal studies; missing response data in phase II cancer clinical trials, such as missing tumor recurrence times when a patient is never examined for tumor recurrence; missing tumor characteristics, such as tumor size, stage, grade, or location; or missing toxicity data. It is also quite common to have missing data on prognostic variables, such as missing estrogen receptor status, nodal status, or size of tumor in cancer studies; CD4 counts in AIDS studies; biomarker variables, such as α-fetoprotein in liver cancer studies; phenotype or genotype information in genetic studies; blood sample measurements (hemoglobin, cholesterol, and so on) in laboratory studies; and so forth. It is important to distinguish between the impact of missing outcome data and that of missing covariate data in an analysis, because their effects can be different depending on the goal of the study. In this article, we characterize such differences in Classifications of Missing Data, once we introduce the various classifications of missing data.

In most analyses appearing in the medical literature, the most common way of dealing with missing (covariate or response) data is to just omit those participants who have any missing data. Such an analysis is called a complete case (CC) analysis. Using a CC analysis has been quite popular, because it is the default analysis for most standard statistical software. For example, in carrying out a Cox regression analysis in a popular statistical package such as SAS (SAS Institute, Cary, NC), STATA (STATA, College Station, TX), or R (http://www.r-project.org), any participant with at least one missing covariate value is omitted from the analysis. It is not uncommon in many of these analyses that 30% of the participants have missing data. This same default

analysis is carried out for missing response data. In many situations, the CC analysis is not an appropriate way to proceed and may lead to inappropriate conclusions, as we will demonstrate in three examples.[3] In this article, we address the various issues in analyzing studies with missing data. We discuss the various classifications of missing data that may arise in a study and demonstrate in several situations that a CC analysis may lead to incorrect results and conclusions as well as to a loss of power in the assessment of treatment effects and prognostic factors. We first discuss the various missing data classifications in the next section.

CLASSIFICATIONS OF MISSING DATA

To carry out statistical inference in the presence of missing data, assumptions are needed regarding the process that generated the missing data, which is called the missing data mechanism. Reasons for missingness include lost data, patient case report forms are incorrectly filled out, personnel error, patient refusal, patient too ill to come to the clinic, study dropout, faulty measurement device, limitations of measurement device, data not entered or updated, and so forth. Valid statistical inference for a particular study requires knowing or estimating the mechanism that generated the missing data. In summary, there are essentially three classifications for missing data mechanisms. These are now listed formally as follows.

Missing Completely at Random

Data are said to be missing completely at random (MCAR) if the failure to observe a value does not depend on any data, either observed or missing. Simple examples of MCAR include lost data, accidental omission of an answer on a questionnaire, accidental breaking of laboratory instrument, and personnel error. In a logistic regression, for example, suppose that the response is completely observed for all participants, whereas some of the covariates are missing for some participants. Then the missing covariate values are MCAR if the chance (probability) of observing the missing covariate is independent of the response as well as independent of the values of the covariates that are fully observed or the covariates that would have been observed (ie, the missing covariates). Under MCAR, the observed data are just a random sample of all the data. A CC analysis may result in larger SEs in the model parameter estimates (ie, loss of efficiency) in this setting, but no bias in the model parameter estimates is introduced when the data are MCAR. Bias is defined as the average difference between model parameter estimates and their true values.

Missing at Random

Data are said to be missing at random (MAR) if, given the observed data, the failure to observe a value does not depend on the data that are unobserved. For example, in cancer clinical trials, information on the size of a primary tumor is often missing, and the size of the primary tumor may depend on the type of the primary tumor, which is often fully observed. If the probability of primary tumor size being missing only depends on the type of primary tumor, then the missingness is considered to be MAR.[6,7] For a more general example involving missing covariates, suppose that the response is completely observed, whereas some covariates may be missing for some participants. The missing values of the covariates are MAR if, given the observed data, the probability of observing the missing covariate is independent of the values of the missing covariate that would have been observed, but this probability is not necessarily independent of the response or the fully observed covariates. MAR is a more realistic assumption than MCAR, but in this case, adjustments must be made, because the observed covariates are no longer a random sample. Clearly, if the missing data are MCAR, then they are MAR. In most MAR scenarios, a CC analysis will be both inefficient and biased. In data that are MAR, if missingness depends only on the fully observed covariates and not on the response, then a CC analysis will lead to unbiased estimates. However, if the missingness depends on the response variable (and not necessarily on the fully observed covariates), then a CC analysis will result in biased parameter estimates.[3] Missing data that are MAR or MCAR, along with the assumption that the parameters of the missing data mechanism are distinct from the parameters of the sampling model (ie, the joint distribution of the covariates and the response), are said to be ignorably missing. In these cases, the missing data mechanism can be ignored in making inferences about the parameters of the sampling model.

Missing Not at Random

The missing data mechanism is said to be nonignorable, or missing not at random (MNAR), if the failure to observe a value depends on the value that would have been observed or other missing values in the data set. MNAR data are most common in longitudinal studies in which missingness is the result of study dropout, toxicity, or illness. For example, the likelihood of obtaining a coping score on a patient in a QOL study often depends on the coping score that would have been observed; patients who are too sick because of toxicity of treatment may not come to the clinic to fill out the QOL questionnaire. Another possibility is that the patient refuses therapy because of toxicity and/or his/her physical condition. A CC analysis in the MNAR setting also leads to biased and inefficient parameter estimates. Valid inferences for MNAR generally require specifying the correct model for the missing data mechanism, distributional assumptions for the response, or both. The resulting parameter estimates and statistics for testing hypotheses may be sensitive to these assumptions. Little et al[3] provide an excellent discussion and examples of the various classifications of missing data.

It is worth mentioning that unfortunately, one cannot determine whether missingness is MNAR or MAR solely based on the data at hand. For this reason, there is a growing consensus among statisticians studying missing data methods that a key component of an analysis is to carry out sensitivity analyses by fitting different missing data mechanisms to examine how sensitive the results are to the assumptions of whether missingness is MNAR.

STATISTICAL METHODS FOR MISSING DATA

The appropriate statistical method depends on the type of missing data mechanism that governs the missingness. Guidelines on missing data in confirmatory clinical trials are under development (http://www.ema.europa.eu/). In this section, we discuss the four most commonly used methods for handling missing data and statistical issues regarding missing outcomes versus covariates.

Maximum Likelihood

A large class of model-based procedures arises from defining a model for the variables with missing values and making statistical

inferences based on what are called maximum likelihood (ML) methods. Model-based methods are quite flexible and clearly set forth underlying model assumptions so that they can be evaluated. In addition, SEs can be easily obtained based on the model using appropriate algorithms and techniques,[8] which take into account the missing data. The literature on ML methods for missing data is enormous, and the articles are too numerous to list here. For example, the literature on missing covariates alone is considerable.[9-18] We refer the reader to the review article by Ibrahim et al[19] and the many references therein. For longitudinal data, we refer the reader to the review article by Ibrahim et al[20] and the many references therein. For survival analysis, we refer the reader to the articles by Chen et al[21] and Herring et al.[22] Modeling issues and sensitivity analyses are addressed in several articles.[2,14,15,23-25]

Multiple Imputation

Multiple imputation (MI) has emerged as a popular technique for dealing with missing data problems. The technique of multiple imputation involves creating multiple complete data sets by filling in values for the missing data. Then, each filled-in data set is analyzed as if it were a complete data set.[3,26] The inferences for the filled-in data sets are then combined into one result by averaging over the filled-in data sets. Many articles[3,26-28] have described MI and some of its variants for missing covariates. It is important to mention here that the imputed values are random; therefore, this randomness must be captured in the SEs of the parameter estimates. The basic idea behind MI, along with parameter estimation and SEs, is described in the Appendix (online only).

There are two types of MI techniques, which are called improper and proper imputation.[29] Improper imputation uses an imputation model that is different from the analysis model, whereas in proper imputation, the imputation model is based on the analysis model. For example, for improper MI, the imputation may be carried out using a normal linear model, and the analysis of the filled-in data may proceed by a logistic regression.[3] Improper MI yields biased estimators. Proper MI, although computationally more intensive, yields unbiased estimates with good large sample properties.[3,26] Finally, we mention that the motivation and basis of proper MI is Bayesian, but the idea of MI itself is quite general and can be applied to other methods, such as hot deck imputation.[3]

Fully Bayesian

Fully Bayesian (FB) methods for missing data (covariate and/or response data) involve specifying distributions for all of the parameters (called prior distributions) as well as specifying distributions for the missing data. The missing data are then imputed from these distributions.[19,30-32] Bayesian methods can easily accommodate missing data without requiring new techniques for statistical inference. In this sense, FB methods are perhaps the most powerful and most general methods for dealing with missing data. ML and MI both have Bayesian connections. The close connections between the ML, MI, and FB procedures and details of the FB methodology are discussed in more detail in the article by Ibrahim et al.[19]

Weighted Estimating Equations

A variety of other approaches requiring fewer model assumptions have been developed to account for missing observations. A general approach called weighted estimating equations (WEEs) has been proposed by Robins et al.[33] WEE methods are the missing data counterparts of what are called generalized estimating equations (GEEs; ie, WEEs are GEE methods adapted to the presence of missing data). General weighted estimating equations[34] are often called doubly robust in the sense that to obtain an unbiased estimate of the regression parameters, either the missing data mechanism or the estimating equations for the missing data given the observed data must be correctly specified, but not both. WEE methods for missing covariates have been discussed in many articles.[16,35-40] We mention here that GEEs are indeed a valid procedure when the missing data are MCAR, but they are generally biased if missingness is MAR or MNAR.

Once a formal statistical model is formulated, estimation can be performed, for example, using any of the four formal methods mentioned, which is a far superior way to handle missing data compared with the ad-hoc CC method. These formal methods, however, can be computationally intensive and may not be available in standard statistical packages, such as SAS, STATA, S-Plus (version 3.3; Statistical Sciences, Seattle, WA), or R. The ML method is often viewed as the gold standard in model fitting. ML methods for generalized linear models are available in some statistical packages, and various versions of MI are available in SAS as well as other packages. FB methods are available in both SAS and WinBUGS (Medical Research Council Biostatistics Unit, Cambridge, United Kingdom); MI, FB, and ML methods often produce similar results in many settings. More details about the software are provided in Software section.

Missing Outcomes Versus Covariates

The impact of missing covariates versus that of missing outcomes on the estimates in a regression model may be quite different. First, we mention that if the data set only has missing responses, and the missing responses are assumed to be MAR, then a CC analysis will lead to unbiased estimates of the regression coefficients (ie, a CC analysis and an ML analysis are equivalent in this case). However, when the responses are MNAR, then the CC analysis as well as the ML analysis assuming MAR will lead to biased estimates. With missing covariates, the story is different. When we have only missing covariates in a regression analysis (no missing responses), then the ML analyses either assuming MAR or MNAR are superior to the CC analysis in terms of bias and efficiency of the parameter estimates. When we have both missing responses and covariates in a data set, as is common in longitudinal studies, then an ML analysis either assuming MAR or MNAR is superior to the CC analysis in terms of bias and efficiency as long as the sampling model and the missing data mechanism are assumed to be correct or approximately correct. If the sampling model and/or the missing data are badly misspecified in an ML analysis, then the ML method can yield poor results. We also refer the reader to the review articles by Ibrahim et al[19,20] for extensive discussions and references on missing data in generalized linear models and models for longitudinal data. We mention here that in analyzing time-to-event data in a randomized clinical trial, one typically examines the treatment effect without adjusting for covariates. This is valid in cases where the censoring mechanism for the time to event only depends on the treatment group and does not depend on other covariates. When the censoring mechanism depends on other covariates, such as age, sex, and so on, then the assessment of the treatment effect should account for these covariates via a Cox regression in time-to-event studies, for example. There are other simple adjustments in some specific scenarios (eg, using propensity score methods for adjusting

for covariates in the assessment of treatment effects[41,42] and the missing data indicator [MDI] method to adjust for partially missing baseline measurements[43]).

SOFTWARE

Existing commercial software for dealing with missing data is still quite limited, although there have been significant advances in software in the past 3 years. The recent article by Horton et al[44] provides an excellent detailed summary of existing software for missing data. The SAS package has the PROC MI procedure for implementing the multiple imputation technique for missing data. However, the PROC MI procedure is best suited for normally distributed data and may not work well with right-censored survival data, discrete responses, or covariate data that are not continuous and normally distributed. Thus, the PROC MI procedure may not be well suited for carrying out Cox regression with missing covariate data or a longitudinal analysis with missing discrete outcomes and/or covariates, for example. Other statistical packages that carry out MI using a variety of techniques include S-Plus, SOLAS (Statistical Solutions, Saugus, MA), Amelia II (http://gking.harvard.edu/amelia), Hmisc (http://www.inside-r.org/packages/cran/Hmisc), ICE (within STATA), IVEware (http://www.isr.umich.edu/src/smp/ive), and MICE (http://mice-software.com). All of these packages have been discussed in detail; we refer the reader to the article by Horton et al[44] for an excellent discussion. The S-Plus package software developed by Schafer[45] uses the ML method for missing response data for categorical or normally distributed responses. S-Plus is also quite limited as far as fitting models with missing data in Cox regression, longitudinal data, and covariate data are concerned. The STATA package has some built-in ad-hoc imputation techniques for handling missing covariate data for normally distributed response data. More recently, the software package XMISS in the LogXact module (Cytel, Cambridge, MA) carries out the ML method for MAR categorical covariates in generalized linear models.[12] In the setting with MAR categorical covariates, PROC MI and XMISS may yield different results. WinBUGS[46] is also quite powerful for implementing the FB method in missing data settings. WinBUGS can essentially handle any type of missing data problem, including missing MAR or MNAR covariates and/or responses in the Cox model, models for longitudinal data, or generalized linear models. When using appropriate prior distributions, WinBUGS yields results similar to those of ML and MI.

MISSING DATA IN COX REGRESSION AND LONGITUDINAL STUDIES

In this section, we present some examples illustrating the importance of incorporating participants with missing data in an analysis. We will compare inferences based on CC analyses with those of the ML method for three common settings in clinical trials: phase II time-to-event clinical trial with MAR covariate (biomarker) data; phase III time-to-event clinical trial with MAR (or MNAR) baseline covariate data; and phase III clinical trial with MAR or MNAR longitudinal response (QOL) data (Appendix; Appendix Tables A1 and A2, online only). In each example, we will show how misleading conclusions can be obtained when a CC analysis is carried out.

Table 1. Parameter Estimates and *P* Values for Liver Cancer Data

Effect	Estimate	SE	*P*
α-fetoprotein			
CC	0.221	0.234	.344
MDI	0.255	0.150	.090
ML	0.392	0.155	.012
γ-globulin			
CC	0.516	0.233	.027
MDI	0.538	0.195	.006
ML	0.536	0.154	.001
Jaundice			
CC	−0.092	0.256	.720
MDI	−0.243	0.155	.116
ML	−0.162	0.155	.295

Abbreviations: CC, complete case; MDI, missing data indicator; ML, maximum likelihood.

Example 1: Liver Cancer Data

We consider a liver cancer data set including 190 patients from two Eastern Cooperative Oncology Group clinical trials to evaluate new treatments in patients with primary liver cancer.[47,48] We are primarily interested in how survival time from study entry to death differs with respect to three baseline characteristics. These three baseline characteristics are associated jaundice (yes or no) and two biochemical markers α-fetoprotein and γ-globulin, each classified as normal or abnormal. Patients with abnormal biochemical markers and/or jaundice are all expected to have shorter survival. Jaundice is always observed, but of the 190 patients in the data set, 109 (57%) are

Table 2. Summary of Small-Cell Lung Cancer Data

Factor	No.
Completely observed variables	
Treatment, frequency	
A	114
B	116
Sex, frequency	
Male	144
Female	86
Age, years	
Mean	62.24
SD	10.17
Time to progression, frequency	
Censored	83
Relapsed	147
Missing covariates	
Apex, frequency	
0	155
1	10
Missing	65
QOL FACT-G score	
Mean	78.14
SD	15.31
Missing	81
Both covariates missing	27
One covariate missing	92

Abbreviations: FACT-G, Functional Assessment of Cancer Therapy–General; QOL, quality of life; SD, standard deviation.

missing α-fetoprotein and/or γ-globulin, of whom 15 (8%) are missing α-fetoprotein, 102 (54%) are missing γ-globulin, and eight (4%) are missing both. The biochemical markers, which are not always easy to obtain, are the covariates with missing values. In the univariate analysis, the estimate, SE, and *P* value for jaundice are −0.218, 0.153, and 0.152, respectively. However, jaundice is highly significant in predicting the missingness of γ-globulin in a multiple logistic regression model for the missing indicator. The adjusted odds ratio, 95% CI, and *P* value of observing γ-globulin are 3.213, 1.723 to 5.989, and < .001 for jaundice, respectively. Because the missing data indicator does depend on jaundice, the assumption of MCAR does not hold for the data. With 57% missing data, a CC analysis using the 81 participants with no missing data could produce highly biased or inefficient estimates, because MAR seems tenable, but MCAR does not, for these data. Table 1 lists the estimates of the parameters for the Cox regression model using the CC, MDI, and ML methods (ML method assumes MAR). The ML method yields different conclusions than the CC, whereas the MDI method is much more consistent with the ML than the CC method. In fact, in terms of *P* values and estimates, the MDI method produces results similar to those of the ML method and yields conclusions similar to those of the ML method. Regarding the effects of the two biochemical markers, γ-globulin is significant ($\alpha = 0.05$) using all three methods (ML, MDI, CC), but α-fetoprotein is significant ($\alpha = 0.05$) under the ML method, marginally not significant under MDI, and not significant under the CC method. Also, the SE seems to be greatly reduced with the ML and MDI methods in general as compared with the CC method. For example, the estimated effect of γ-globulin is approximately the same using the CC and ML methods, but the estimated SE is reduced from 0.233 for CC to 0.155 for the ML method, which is a 34% reduction. The MDI estimate corresponding to the unknown category of α-fetoprotein is 0.959, and the MDI estimate corresponding to the unknown category of γ-globulin is −0.01. These results explain why the ML estimate of α-fetoprotein is larger than the MDI estimate, and the ML and MDI estimates are similar for γ-globulin, because the MDI method treats the unknown category as the reference group in this example. For this example, the ML and MDI methods produce similar conclusions. To examine the MAR assumption under ML, we carried out several sensitivity analyses by fitting several MNAR models and found that the estimates and *P* values under these MNAR models were similar to the estimates under the MAR model. In addition, the *P* values corresponding to the two missing covariates in the logistic regression models for the missing indicators are greater than .14, suggesting that there is no statistically significant evidence against the MAR assumption for these data. For additional details on the modeling and analyses of these data, we refer to reader to the work of Lipsitz et al[14] and Lipsitz and Ibrahim.[1]

Table 3. Parameter Estimates and *P* Values for Small-Cell Lung Cancer Data

Effect	Full Model			Reduced Model		
	Estimate	SE	*P*	Estimate	SE	*P*
Treatment						
CC	0.471	0.253	.062	0.383	0.193	.047
MDI	0.469	0.174	.007	0.486	0.168	.004
MAR	0.477	0.175	.006	0.491	0.169	.004
MNAR	0.478	0.175	.006	0.491	0.169	.004
Sex						
CC	0.068	0.243	.780			
MDI	0.177	0.178	.320			
MAR	0.174	0.180	.334			
MNAR	0.174	0.180	.336			
Age						
CC	−0.020	0.130	.878			
MDI	−0.023	0.088	.795			
MAR	−0.021	0.090	.812			
MNAR	−0.021	0.090	.814			
Apex						
CC	0.879	0.411	.032	0.804	0.371	.030
MDI	0.862	0.372	.021	0.777	0.366	.034
MAR	0.914	0.381	.016	0.810	0.374	.030
MNAR	0.923	0.380	.015	0.808	0.372	.030
QOL FACT-G score						
CC	−0.138	0.119	.247			
MDI	−0.065	0.100	.518			
MAR	−0.052	0.105	.624			
MNAR	−0.051	0.106	.628			

NOTE. If we fit treatment alone to the data with 230 observations, the estimate, SE, and *P* value are 0.482, 0.168, and .004, respectively; if we fit treatment adjusting for sex, the estimate, SE, and *P* value for treatment are 0.477, 0.168, and .005, respectively; if we fit treatment adjusting for sex and age, the estimate, SE, and *P* value for treatment are 0.465, 0.174, and .008, respectively; and if we fit treatment adjusting for sex, age, and apex, the estimates, SEs, and *P* values for treatment are 0.472, 0.175, and .007 under MAR and 0.473, 0.175, and .007 under MNAR, respectively. In addition, we add the QOL FACT-G score to the reduced model in the table. Under this three-covariate model, the estimates, SEs, and *P* values for treatment are 0.486, 0.234, and .038 under CC, 0.484, 0.168, and .004 under MDI, and 0.489, 0.169, and .004 under MAR and MNAR, respectively; the estimates, SEs, and *P* values for apex are 0.863, 0.407, and .034 under CC, 0.819, 0.369, and .027 under MDI, 0.853, 0.380, and .025 under MAR, and 0.863, 0.380, and .023 under MNAR, respectively; and the estimates, SEs, and *P* values for QOL FACT-G score are −0.146, 0.115, and .204 under CC, −0.084, 0.096, and .381 under MDI, −0.062, 0.100, and .531 under MAR, and −0.061, and 0.100, and .542 under MNAR, respectively.

Abbreviations: CC, complete case; FACT-G, Functional Assessment of Cancer Therapy–General; MAR, missing at random; MDI, missing data indicator; MNAR, missing not at random; QOL, quality of life.

Example 2: Phase III Advanced Non–Small-Cell Lung Cancer Clinical Trial

We consider data from a phase III clinical trial of advanced non–small-cell lung cancer conducted by the University of North Carolina at Chapel Hill (LCCC 9719). The results of this study were reported by Socinski et al.[49] The goal of this trial was to compare a

Table 4. Minimal Information for Handling Missing Data

Option	Minimal Information
1	Summary of missing covariate and/or response data, including fractions of missing data
2	Summary of how missing values are analyzed, including: (1) CC analysis as benchmark analysis; (2) logistic regression analysis of missing indicators using completely observed variables as covariates; (3) if relevant, summary of models and MAR analysis results using either ML, MI, FB, or WEE methods; (4) if relevant, summary of MNAR models and MNAR analysis using either ML, MI, FB, or WEE methods; (5) if relevant, sensitivity analysis results for MNAR; and (6) if relevant, comparison and summary of results of CC, MDI, MAR, and MNAR

Abbreviations: CC, complete case; FB, fully Bayesian; MI, multiple imputation; ML, maximum likelihood; MAR, missing at random; MDI, missing data indicator; MNAR, missing not at random; WEE, weighted estimating equation.

defined duration of therapy (arm A) with continuous therapy followed by second-line therapy (arm B) to determine optimal duration of therapy in patients with non–small-cell lung cancer. LCCC 9719 included 230 patients. We consider here five prognostic factors: treatment (two arms: A and B, coded as 1 and 0), sex (female and male, coded as 0 and 1), age in years, apex (two levels: 0 and 1, where 1 indicates that the tumor was at the top of the lung), and baseline QOL Functional Assessment of Cancer Therapy–General (FACT-G) score. For these five prognostic factors, apex and QOL FACT-G score had missing information, whereas treatment, sex, and age were completely observed for all patients. In this data set, there is a total missing covariate data fraction of 51.74% on these two covariates. The outcome variable is time to progression in months, which is continuous and subject to right censoring. The median follow-up time is 3.94 months (range, 0.10 to 27.61 months). A summary of the data set is provided in Table 2.

We fit the Cox regression model to the LCCC 9719 data with these five covariates. We carry out four analyses (ie, CC, MDI, MAR covariates, and MNAR covariates). The missing data mechanism for the model assuming MNAR is a logistic regression model with survival time, missing data indicator, apex, treatment, sex, age, and QOL as covariates in the model. There are no interaction terms in the missing data mechanism. Table 3 lists ML estimates along with SEs and *P* values for the five covariates. In the CC analysis, the *P* value for treatment is .036 if we fit the treatment covariate alone; if we fit the treatment covariate along with sex, age, apex, and QOL, the *P* value for treatment is .062, whereas the *P* value for apex is .032. In the univariate Cox regression analysis, sex and age are not significant, and the corresponding *P* values are 0.332 and 0.433, respectively. However, sex is significant in the logistic regression model for the missing indicator of apex; the odds ratio and *P* value are 0.49 and .033, respectively. Also, an older patient tends to have a worse QOL FACT-G score ($P = 0.006$). Thus, it is of clinical importance to include both of these covariates in the analysis of these data. In the MDI analysis, the *P* value for treatment is .007, whereas the *P* value for apex is .021. Also, the *P* values are .006 and .006 for treatment and .016 and .015 for apex under the MAR and MNAR models, respectively. We also consider a reduced model with treatment and apex only. Under this reduced model, the *P* values for treatment are .047, .004, .004, and .004 corresponding to CC, MDI, MAR, and MNAR, respectively, whereas the *P* values for apex are .030, .034, .030, and .030 corresponding to CC, MDI, MAR, and MNAR, respectively. These results imply that under all three methods (MDI, MAR, MNAR), treatment and apex are significantly associated with time to disease progression at the 5% significance level. Thus, continuous therapy followed by second-line therapy may have a strong effect (ie, more beneficial) compared with defined duration of therapy with respect to time to progression based on an analysis using all patients. Also, the SEs based on the ML method (MAR or MNAR) are consistently smaller than those from the CC analysis for all of the covariates.

In addition, from Table 3, we see that both sets of ML estimates of the regression coefficients for all covariates are similar under the MAR and MNAR scenarios. Because the MAR model is a special case (ie, a submodel) of the MNAR model, these results suggest that there is no evidence against the MAR assumption in the LCCC 9719 data. Moreover, we note that the MDI, MAR, and MNAR methods all yield similar conclusions about the effects of the covariates. This example shows the importance of using all of the participants whenever possible in carrying out an analysis. For more details on the modeling and analysis of these data, we refer the reader to the article by Chen et al.[50]

CONCLUSION

We see from these examples how statistical analyses may be affected if participants with missing values are omitted from an analysis. Misleading results might be obtained regarding the effect of treatment, unreliable *P* values may be obtained, and assessments of the importance of prognostic factors may be inaccurate. The MAR results for examples 1 and 2 can be obtained using PROC MI in SAS or WinBUGS. The MNAR results can be obtained only using WinBUGS. In addition, the MDI results are easily obtained in any statistical package that carries out Cox regression; this method is much easier to implement than the ML method, because it does not require special tools or new statistical methods for model fitting. Table 4 provides a list of the minimal information that should be included in the methods section of an article describing how missing data should be handled. In general, it is strongly recommended that one avoid performing a CC analysis whenever possible; one should use ML, MI, FB, or WEE methods as appropriate. Finally, it is important to note here that the ML (MAR or MNAR) and MDI methods may yield imprecise estimates and *P* values, and hence wrong conclusions, if the missing data fraction is too high in a given data set. As a general rule of thumb, if the missing data fraction for a given variable (covariate or outcome) is greater than 50%, then one may obtain imprecise estimates and *P* values in the regression model. In such cases, one must conduct these analyses with great caution and perform several sensitivity analyses.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The author(s) indicated no potential conflicts of interest.

AUTHOR CONTRIBUTIONS

Conception and design: Joseph G. Ibrahim
Financial support: Joseph G. Ibrahim
Administrative support: Joseph G. Ibrahim
Provision of study materials or patients: Joseph G. Ibrahim
Collection and assembly of data: Joseph G. Ibrahim
Data analysis and interpretation: Joseph Ibrahim, Ming-Hui Chen
Manuscript writing: All authors
Final approval of manuscript: All authors

REFERENCES

1. Lipsitz SR, Ibrahim JG: Estimating equations with incomplete categorical covariates in the Cox model. Biometrics 54:1002-1013, 1998
2. Ibrahim JG, Chen MH, Lipsitz SR: Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. Biometrika 88:551-564, 2001
3. Little RJA, Rubin DB: Statistical Analysis With Missing Data (ed 2). Hoboken, NJ, John Wiley and Sons, 2002
4. Verbeke G, Molenberghs G: Linear Mixed Models for Longitudinal Data. New York, NY, Springer, 2000
5. Molenberghs G, Verbeke G: Models for Discrete Longitudinal Data. New York, NY, Springer, 2005

6. Kirkwood JM, Strawderman MH, Ernstoff MS, et al: Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: The Eastern Cooperative Oncology Group Trial EST 1684. J Clin Oncol 14:7-17, 1996

7. Kirkwood JM, Ibrahim JG, Sondak VK, et al: High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of Intergroup Trial E1690/S9111/C9190. J Clin Oncol 18:2444-2458, 2000

8. Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via the EM algorithm. JR Stat Soc B (Methodological) 39:1-38, 1977

9. Chen T, Fienberg SE: The analysis of contingency tables with incompletely classified data. Biometrics 32:133-144, 1976

10. Fuchs C: Maximum likelihood estimation and model selection in contingency tables with missing data. J Am Stat Assoc 77:270-278, 1982

11. Little RJ, Schluchter MD: Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika 72:497-512, 1985

12. Ibrahim JG: Incomplete data in generalized linear models. J Am Stat Assoc 85:765-769, 1990

13. Vach W: Logistic Regression With Missing Values in the Covariates. New York, NY, Springer-Verlag, 1994

14. Lipsitz SR, Ibrahim JG: Using the EM-algorithm for survival data with incomplete categorical covariates. Lifetime Data Anal 2:5-14, 1996

15. Ibrahim JG, Lipsitz SR, Chen MH: Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. JR Stat Soc B (Statistical Methodology) 61:173-190, 1999

16. Lipsitz SR, Ibrahim JG, Zhao LP: A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. J Am Stat Assoc 94:1147-1160, 1999

17. Lipsitz SR, Ibrahim JG, Chen MH, et al: Non-ignorable missing covariates in generalized linear models. Stat Med 18:2435-2448, 1999

18. Ibrahim JG, Lipsitz SR, Horton N: Using auxiliary data for parameter estimation with non-ignorably missing outcomes. JR Stat Soc C (Applied Statistics) 50:361-373, 2001

19. Ibrahim JG, Chen MH, Lipsitz SR, et al: Missing-data methods for generalized linear models: A comparative review. J Am Stat Assoc 100:332-346, 2005

20. Ibrahim J, Molenberghs G: Missing data methods in longitudinal studies: A review. Test (Madr) 18:1-43, 2009

21. Chen HY, Little R: A test of missing completely at random for generalised estimating equations with missing data. Biometrika 86:1-13, 1999

22. Herring AH, Ibrahim JG: Likelihood-based methods for missing covariates in the Cox proportional hazards model. J Am Stat Assoc 96:292-302, 2001

23. Lipsitz SR, Ibrahim JG: A conditional model for incomplete covariates in parametric regression models. Biometrika 83:916-922, 1996

24. Ibrahim JG, Lipsitz SR: Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. Biometrics 52:1071-1078, 1996

25. Ibrahim JG, Chen MH, Lipsitz SR: Bayesian methods for generalized linear models with covariates missing at random. Can J Stat 30:55-78, 2002

26. Rubin DB: Multiple Imputation for Nonresponse in Surveys. New York, NY, Wiley, 1987

27. Gelman A, Carlin JB, Stern HS, et al: Bayesian Data Analysis (ed 2). Boca Raton, FL, Chapman and Hall/CRC Press, 2004

28. Raghunathan TE, Lepkowski JM, Van Hoewyk J, et al: A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv Methodol 27:85-96, 2001

29. Nielsen SF: Proper and improper multiple imputation. Int Stat Rev 71:593-607, 2003

30. Scharfstein DO, Halloran ME, Chu H, et al: On estimation of vaccine efficacy using validation samples with selection bias. Biostatistics 7:615-629, 2006

31. Chu H, Halloran ME: Estimating vaccine efficacy using auxiliary outcome data and a small validation sample. Stat Med 23:2697-2711, 2004

32. Chu H, Chen SN, Louis TA: Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. J Am Stat Assoc 104:512-523, 2009

33. Robins JM, Rotnitzky A, Zhao LP: Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89:846-866, 1994

34. Robins JM, Ritov Y: Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. Stat Med 16:285-319, 1997

35. Robins JM, Rotnitzky A, Zhao LP: Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 90:106-121, 1995

36. Robins JM, Rotnitzky A: Semiparametric efficiency in multivariate regression models with missing data. J Am Stat Assoc 90:122-129, 1995

37. Rotnitzky A, Robins JM: Semiparametric regression estimation in the presence of dependent censoring. Biometrika 82:805-820, 1995

38. Rotnitzky A, Robins JM, Scharfstein DO: Semiparametric regression for repeated outcomes with nonignorable nonresponse. J Am Stat Assoc 93:1321-1339, 1998

39. Scharfstein DO, Rotnitzky A, Robins JM: Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Am Stat Assoc 94:1096-1120, 1999

40. Scharfstein DO, Irizarry RA: Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. Biometrics 59:601-613, 2003

41. Baker SG, Fitzmaurice GM, Freedman LS, et al: Simple adjustments for randomized trials with nonrandomly missing or censored outcomes arising from informative covariates. Biostatistics 7:29-40, 2006

42. Baker S, Freedman L: A simple method for analyzing data from a randomized trial with a missing binary outcome. BMC Med Res Methodol 3:8, 2003

43. White IR, Thompson SG: Adjusting for partially missing baseline measurements in randomized trials. Stat Med 24:993-1007, 2005

44. Horton NJ, Kleinman KP: Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat 61:79-90, 2007

45. Schafer JL: Analysis of Incomplete Multivariate Data. New York, NY, Chapman and Hall/CRC, 1997

46. Spiegelhalter DJ, Thomas A, Best NG: WinBUGS 1.4.1: Bayesian inference using Gibbs sampling. http://www.mrc-bsu.cam.ac.uk/bugs

47. Falkson G, Cnaan A, Simson IW, et al: A randomized phase II study of acivicin and 4′deoxydoxorubicin in patients with hepatocellular carcinoma in an Eastern Cooperative Oncology Group study. Am J Clin Oncol 13:510-515, 1990

48. Falkson G, Lipsitz S, Borden E, et al: Hepatocellular carcinoma: An ECOG randomized phase II study of interferon-beta and menagoril. Am J Clin Oncol 18:287-292, 1995

49. Socinski MA, Schell MJ, Peterman A, et al: Phase III trial comparing defined duration of therapy versus continuous therapy followed by second-line therapy in advanced-stage IIIB/IV non–small-cell lung cancer. J Clin Oncol 20:1335-1343, 2002

50. Chen MH, Ibrahim JG, Shao QM: Maximum likelihood inference for the Cox regression model with applications to missing covariates. J Multivar Anal 100:2018-2030, 2009

