American Journal of
EPIDEMIOLOGY

Volume 142
Number 12
December 15, 1995

Sponsored by the Society for Epidemiologic Research

REVIEWS AND COMMENTARY

A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses

Sander Greenland[1] and William D. Finkle[2]

Epidemiologic studies often encounter missing covariate values. While simple methods such as stratification on missing-data status, conditional-mean imputation, and complete-subject analysis are commonly employed for handling this problem, several studies have shown that these methods can be biased under reasonable circumstances. The authors review these results in the context of logistic regression and present simulation experiments showing the limitations of the methods. The method based on missing-data indicators can exhibit severe bias even when the data are missing completely at random, and regression (conditional-mean) imputation can be inordinately sensitive to model misspecification. Even complete-subject analysis can outperform these methods. More sophisticated methods, such as maximum likelihood, multiple imputation, and weighted estimating equations, have been given extensive attention in the statistics literature. While these methods are superior to simple methods, they are not commonly used in epidemiology, no doubt due to their complexity and the lack of packaged software to apply these methods. The authors contrast the results of multiple imputation to simple methods in the analysis of a case-control study of endometrial cancer, and they find a meaningful difference in results for age at menarche. In general, the authors recommend that epidemiologists avoid using the missing-indicator method and use more sophisticated methods whenever a large proportion of data are missing. *Am J Epidemiol* 1995;142:1255–64.

biostatistics; epidemiologic methods; logistic regression; missing data; odds ratio; relative risk

Many, if not most, epidemiologic studies will suffer from missing (unrecorded) values on some variables and subjects. The problems of analyzing data with missing values have been extensively studied in the statistics literature (1–3), but have so far received little attention in epidemiologic textbooks. One book (4) devotes a brief section to regression analysis in the presence of missing covariate (regressor) values. For reasons discussed below and elsewhere (5), however, this treatment is unsatisfactory. Vach and Blettner (5, 6) provide a lucid article on missing data methods for tabular analyses. A more technical review emphasizing normal linear regression has been given by Little (3). A comparative study of basic methods for handling missing data in linear regression has recently been given by Jones (7). We will review these discussions in the context of epidemiologic regression analyses, and provide a small simulation study to illustrate that some common methods for handling missing covariate values in logistic regression are unsatisfactory, and that more sophisticated methods are preferable. We will also contrast the results of several methods in a logistic-regression analysis of a case-control study of endometrial cancer.

Received for publication December 14, 1994, and in final form March 28, 1995.

Abbreviations: CS, complete-subject; MAR, missing-at-random; MCAR, missing completely at random; MI, multiple imputation; MM, modified missing indicator; RI, regression imputation; RMSE, root-mean-squared error.

[1] Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA.
[2] Consolidated Research, Inc., Los Angeles, CA.
Correspondence to Dr. Sander Greenland, Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90095–1772.

MISSING-DATA METHODS

The simplest approach is complete-subject (or complete-case) analysis, in which only subjects with all values recorded for all covariates are retained in the analysis (1–5). One book states that complete-subject analysis "is the only approach that assures no bias is introduced under any circumstances" (4, p. 231). It is well known in the statistics literature, however, that complete-subject analysis can be biased (1–3). Vach and Blettner (5) provide some simple examples to illustrate this point, in which the subjects with complete data are a biased subsample of all subjects. Even when complete-subject analysis is valid, it can be very inefficient, in the sense of producing estimates with higher variance than can be obtained with other equally valid methods (1–3).

Given that complete-subject analysis is inefficient and can be biased, it seems natural to search for better methods. One approach with a long history is to use the complete subjects in a weighted regression, with weights that vary inversely with the estimated probability of being complete (1, 2, 8). The "corrected complete case" analysis discussed by Vach and Blettner (5) may be viewed as a special case of this approach. Although the basic idea is simple, the weights and standard errors can become extremely complex if the data and missing-data patterns are not simple (8).

Another simple idea developed before the modern computing era is the indicator method (4, 7, 9). For each variable with missing values, one creates a missing-value indicator to accompany the variable in all analyses. This missing-value indicator takes the value 1 wherever the original variable is missing, 0 elsewhere. For example, if X_1 is age at menarche and some subjects have X_1 missing, one creates a missing-age-at-menarche indicator M_1 such that $M_1 = 1$ among subjects with X_1 missing, $M_1 = 0$ among the rest. One then replaces age at menarche in the regression with the missingness indicator, M_1, and the product of age at menarche and one minus the indicator, $X_1(1 - M_1)$; note that the latter product is zero if X_1 is missing. The indicator method can yield estimates with much reduced standard errors relative to the complete-subject method. Unfortunately, the method is also biased under most conditions (e.g., in linear regression (7)). The degree of bias depends heavily on whether X_1 is the primary study variable or simply a confounder, and the degree of confounding by X_1 if X_1 is left uncontrolled (4).

When only a single categorical covariate has missing values, the indicator method is equivalent to creating an additional "missing" category for the covariate. Vach and Blettner (5) provide examples illustrating how this approach can lead to appreciable bias if the covariate is an important confounder of the study effect: Even if the $M_1 = 1$ (X_1 missing) stratum is just a random sample of all subjects, the stratum will yield a confounded estimate of the exposure effect. This biased estimate will then be averaged with the remaining unconfounded X_1-specific estimates, resulting in a biased summary estimate. Despite these problems, we have discovered through inquiries among epidemiologists that the indicator method is still widely perceived as a formally correct method of handling missing values.

One may reduce the bias of the indicator method by further adding all covariate-indicator products $X_jM_k(1 - M_j)$ to the regression (9), where j ranges over all covariate indices and k ranges over all missing-indicator indices, with $k \neq j$. This modified indicator method is equivalent to the complete-case method in the examples given by Vach and Blettner (5), and so gives the same estimates in these examples. More generally, because of the extra parameters estimated in the modified-indicator method, it is not clear that it can offer any worthwhile efficiency gains over the complete-subject method. This issue will be partially addressed in the simulation examples below.

A somewhat more sophisticated idea is to fill in (impute) missing values for each subject with values predicted from the rest of the data (1–5, 9–12). For example, a subject with a missing value for age at menarche could be assigned a value predicted (imputed) from a linear regression of age at menarche on other covariates, with the latter regression estimated from the complete subjects (1–3). While intuitively appealing, the standard errors of the final coefficients will be underestimated if one simply treats the filled-in data as if it were a data set with no missing values (1–3); correct standard-error formulas can be complex, and special weighting for the regression may be needed (2, 3). Iterative refinements of the weights and imputations can lead naturally to maximum-likelihood methods in important cases, such as contingency tables (2, 12).

Several more advanced and formally justifiable methods have been developed, which offer greater potential for precision and validity than the above basic methods. They may be roughly classified into three groups:

1) Multiple-imputation (MI) methods, which may be viewed as refinements of the prediction-imputation approach described above (1–3, 10, 11). These methods begin by generating multiple copies of the original data set, each with missing values replaced by values randomly generated according to some model for the distribution of incomplete regressors and its dependence on complete regressors and the outcome vari-

able. Each of these imputed data sets is analyzed as if it were complete; the different results from the data sets are then combined in a manner that takes account of the imputation variability (1–3, 10, 11).

2) Maximum-likelihood and maximum pseudo-likelihood methods, in which a joint model for the disease under study, the covariate distribution, and possibly the missing-data process is fit (1–3, 12–19).

3) Weighted estimating equation methods, in which a model for the missing-data process is used to provide special weights and covariates for the disease regression analysis (8, 20). The weighted-regression approach described above falls in this class. These methods may be viewed as extensions of pseudo-likelihood methods for two-stage designs (21) to general missing-data problems (8).

These three groups of advanced methods contain many variants. We will not attempt to describe the variants here, because considerable theory is required to do so, and the methods are not yet available in widely distributed software.

Both the maximum-likelihood and weighted estimating equation methods can achieve the maximum efficiency one can expect under their respective distributional assumptions (2, 8). Multiple imputation methods were developed in part to obtain most of the benefits of these advanced approaches without having to develop special software or to rely on unfamiliar modeling techniques (1, 2). Nonetheless, because of the need to generate imputations, even the simplest versions of MI require considerably more effort than the more basic methods. Thus, it is worth exploring what conditions (if any) render the effort worthwhile.

ASSUMPTIONS FOR THE METHODS

As with all analyses, validity of the above methods depend on absence of other biases (such as those due to measurement error or residual confounding), as well as accuracy of the assumed model for the dependence of the outcome (disease) on covariates. In the following discussion, we will take these standard preconditions as given.

Perhaps the most stringent assumption one can make about missing values is that they are missing completely at random, or MCAR (1–3, 5). This is equivalent to assuming that, for each variable, the observed (nonmissing) values effectively constitute a simple random sample of the values for all study subjects, so that whether one is observed or not for a given variable is independent of any other variable and independent of whether one is observed or not for any other variable. As an example, in the record-based study described below, some data were missing from the records of over half the subjects. If the absence was solely due to random failures of attending personnel to request or note information (which is somewhat plausible, given that the affected variables might seem irrelevant for patient care), then the MCAR assumption would be satisfied.

The MCAR assumption is sometimes generalized to a stratified-MCAR assumption that requires only randomness of missing data within levels of completely observed covariates. As an example, in the study described below, suppose that age at menarche was sometimes missing because the subject failed to recall her exact age at menarche. We might expect this failure to be related to current age, but to be random among women of the same age. If so, and if data on the remaining variables were missing completely at random or in a manner related only to age, we could say that the data were MCAR within levels of age. For simplicity, we will use MCAR to refer to the narrower unstratified MCAR definition, and stratified MCAR to refer to the broader definition.

It is reasonable to demand that a missing-data analysis method give valid estimates under the simple MCAR assumption, provided no other sources of bias are present. With this assumption, complete-subject analysis is equivalent to analysis of a simple random sample of all subjects, and so will be as valid as an analysis of all the subjects would have been (although less precise). Thus, complete-subject analysis meets our basic demand.

Using the MCAR assumption, the weights in the weighted-regression approach become equal, and the resulting weighted-regression analysis is the same as complete-subject analysis. If the MCAR assumption is correct but the weights are not derived using this assumption, the weighted regression estimates can remain consistent, although their standard errors would require special formulas (8). In contrast, the ordinary indicator method does not meet our basic demand: The first example in table 5 of Vach and Blettner (5) is one in which the MCAR assumption is satisfied but the ordinary indicator method gives a biased answer (the column labeled "AC" in their table).

The MCAR assumption is unrealistically stringent for general use. Epidemiologic studies routinely expect completeness of records or questionnaire responses to be related to disease or exposure status (5). A less stringent assumption is that values are missing at random (MAR): Whether a value is missing or not (e.g., whether $M_1 = 1$ or 0) does not depend on any unobserved (missing) covariate or outcome value (1, 2). Thus, under the MAR assumption, the missingness indicator M_1 for a covariate X_1 cannot depend on any unobserved variable values, although it can depend on observed values. For example, the MAR assumption

requires that whether or not X_1 is observed cannot depend on X_2 when X_2 is missing, but can depend on X_2 when X_2 is observed.

One may also define a stratified-MAR condition in which the data are MAR only within joint levels of completely observed covariates and the outcome. This assumption is often confused with the stronger stratified-MCAR assumption. This confusion is understandable, because it is often difficult to imagine realistic scenarios in which the data are stratified-MAR but not stratified-MCAR. For example, it may seem counterintuitive that missingness of X_1 (as represented by M_1) could depend on X_2 when and only when X_2 is observed. The MAR assumption, however, is not intended to be intuitive, but instead provides a minimal condition under which valid effect estimates can be obtained while ignoring (i.e., without having to model) the missing data process (subject, of course, to other validity conditions) (1–3, 5).

As illustrated by Vach and Blettner (5), both the complete-case and indicator methods can be biased when the MAR assumption holds (1–3). The remaining methods will be valid under MAR provided the models used for weighting and imputation are accurate, and imputations are handled appropriately (especially in standard-error calculations) (1–3, 8). The latter assumptions should not be taken lightly, however; in particular, valid methods for handling predictive imputations are not generally available.

The MAR assumption is itself not always reasonable (5, 10). For example, failure to answer a question on sexual preference (heterosexual/homosexual) would likely depend on the true preference, even conditional on observed values. If the MAR assumption fails, none of the above methods can be guaranteed to be valid. Furthermore, the MAR assumption cannot be tested without further data or untestable assumptions if no covariates are available beyond those already in the analysis (5, 8). In this respect, the MAR assumption is much like the no-residual-confounding assumption used in causal inference, which asserts that there is no confounding within levels of controlled covariates; this assumption is also not identifiable (testable) without further data or untestable assumptions.

Unlike the no-residual-confounding assumption, however, the MAR assumption has a certain asymmetric property that limits its a priori plausibility. It allows a covariate with no missing values to predict the observational status (observed/missing) of any other variable in a virtually unrestricted manner, but it forbids any association between the missing values of a covariate and the observational status of any other covariate. For example, given MAR, if both religious preference and abortion history can have missing values, among subjects with missing values of religious preference there can be no association of religious preference with observation of abortion history—even if religious preference when observed *is* associated with observation of abortion history.

The MAR assumption may be acceptable within levels of certain completely recorded covariates, in which case a stratified-MAR assumption can be used subject to control of these covariates. Otherwise, when no MAR assumption is acceptable, one may turn to methods that involve joint estimation of the missing-data process and the distribution of the study variables (nonignorable nonresponse methods (1, 2)). Such methods are difficult to apply, however, and are very sensitive to the missing-data model assumed (1, 2). Therefore, we will not consider these methods here. Instead, we present simulations from both MAR and non-MAR models for the missing-data process, and we examine how the simpler methods described above perform under commonly assumed disease regression models. The results provide some indication of the relative performance among basic missing-data methods in epidemiologic analyses.

SIMULATION EXPERIMENTS

Rationale and design

Because of the relative computational intensity of missing-data methods other than complete-subject and indicator methods, simulation studies outside of normal linear regression have been limited, and critical reviews have tended to limit themselves to a few data or numerical examples (3, 5, 9, 11). While such examples provide useful information about biases, they do not address measures of accuracy (such as mean-squared error) or confidence-interval coverage. To address these issues under the mechanisms and methods considered here, a Monte Carlo simulation study is needed. The following study is only intended to provide an indication of the potential differences among common approaches.

The methods chosen for simulation study were selected because they can be carried out with readily available software, plus possibly some simple calculations. These methods are thus feasible for a typical working epidemiologist, and run with acceptable speed for simulation when several covariates may have missing values:

1) Complete-subject regression (CS).
2) Ordinary missing-indicator regression (OM).
3) Modified missing-indicator regression (MM).
4) Naive regression imputation (RI): Replace missing values with predicted values from MM regressions of each regressor (covariate) on the remaining

regressors. Perform regression on the resulting filled-in data set.

5) Multiple imputation (MI): Fit a joint normal distribution for the regressors given disease to the complete subjects, then create K multiple imputed data sets by replacing missing values with draws from the conditional distribution of the missing values given the observed values.

The multiple data sets produced by MI were analyzed in two different ways:

A) Rubin's method (1, 2): Separately analyze each of the data sets, average the estimates to get the final point estimate, and estimate the variances by the average of the separate estimated variances plus the variance of the separate point estimates.

B) Robins' method (20): Combine all the data sets and analyze as a single weighted data set with correlated observations. Each complete record contributes one independent record to the new data set, while each incomplete record contributes K correlated records to the new data set. The K new records missing from a single incomplete record are treated as having weights and correlations proportional to the amount of information contained in the original incomplete record (20). (This type of approach has been termed fractionally weighted imputation (22).)

The multiple imputations used here are "improper" in Rubin's sense in that they do not involve sampling of the parameters in the imputation (regressor) model (1, 2). This reduces variability in the point estimator relative to proper imputation, and so should (and, in the simulations below, does) lead to somewhat conservative behavior of intervals (supranominal coverage) based on the variance formulas used here when the imputation model is correctly specified.

To speed computations, only $K = 2$ imputations were used for the MI simulations, and this led to estimators with roughly 15–35 percent higher variance than that achievable with a large number of imputations (say, 50) or under maximum likelihood with ignorable missingness (whose variance gives a lower bound to that of MI). Thus, the MI results serve mainly to indicate some non-MAR conditions under which MI becomes unacceptably biased. For efficiency comparisons, the simulations also calculated variance estimates from maximum likelihood assuming MAR and normal covariates (which can be estimated without maximizing the likelihood by calculating the multiple imputation variance for K equals infinity (20)), and from maximum likelihood based on the full data (with no missing values). Following Little (3), the outcome Y (here, a disease indicator, $Y = 1, 0$ for no disease) was excluded from the regression-imputation models, but was included as a regressor in the multiple-imputation models (inclusion of Y in regression imputation produces bias (3)).

Two regressors X_1 and X_2 were generated from a bivariate normal pair Z_1 and Z_2 with zero means. In tables 1–3 below, $X_j = Z_j$ with unit variance. In table 4, $X_j = \exp(Z_j)$, where Z_j has variance $\ln(\tau) \doteq 0.5$ and τ is the "golden ratio" $(1 + \sqrt{5})/2$, so that X_j is lognormal with unit variance. Note that under this lognormal scheme, the correlation r of X_1 and X_2 cannot be below $-1/\tau = -0.62$ (23), hence the first column of table 4 is set to -0.40. One degree-of-freedom chi-squared X_j were also examined but gave qualitatively similar results to the lognormal experiments and so are not presented. Y was generated from X_1 and X_2 via the logistic model

$$P(Y = 1|X_1 = x_1, X_2 = x_2) = \operatorname{expit}(\alpha + \beta_1 x_1 + \beta_2 x_2),$$

where $\operatorname{expit}(z) = e^z/(1 + e^z)$ is the logistic transform. Finally, for each X_j, a missing indicator M_j was generated from X_1, X_2 and Y via

$$P(M_j = 1|X_1 = x_1, X_2 = x_2, Y = y) = \operatorname{expit}(a_j + b_{1j}x_1 + b_{2j}x_2 + d_j y).$$

If both covariates have missing values, MAR requires $b_{1j} = b_{2j} = 0$ for both j, and MCAR further requires $d_j = 0$ for both j. Y was assumed to be always observed.

To conserve space, results from only 12 of over 100 experiments are presented here. Other settings of the simulation parameters (r, α, β_j, a_j, b_{1j}, b_{2j}, d_j, and sample size) produced the same qualitative conclusions. In the experiments presented here, $\beta_1 = \beta_2 = \ln(2)$, so that a unit increase in X_j produced a twofold increase in the odds of $Y = 1$, and $\alpha = -2\ln(2)E(X)$, so that on average $Y = 1$ in half the generated subjects. As one might expect from the estimability of the β_j under outcome-stratified (case-control) sampling, results were completely insensitive to α except for small-sample effects. In all tables, the sample size per trial was 400, which was adequate to avoid such effects for the chosen α.

In tables 2–4, $b_{1j} = b_{2j} = b = \ln(1.2)$, so that a unit increase in X_j produced a 20 percent increase in the odds of missing data in either regressor. In tables 1, 3, and 4, $d_j = d = \ln(4)$, so that subjects with $Y = 1$ had four times the odds of missing data for either regressor, relative to subjects with $Y = 0$. In all tables, the intercepts a_j were set so that on average half the subjects would have no missing data.

In the simulations reported here, X_1 and X_2 are completely exchangeable. Each table column gives

results from 1,000 generated samples (trials), so that each number summarizes results from 2,000 coefficient estimates. This yields very small simulation standard errors for all numbers. For example, the simulation standard error for a confidence-interval coverage percent is approximately $100(0.05(0.95)/2{,}000)^{1/2} = 0.5$ percent when the true coverage probability is 95 percent and the regressors are uncorrelated.

In all the simulations reported below, coverages refer to coverages of nominal 95 percent confidence intervals computed from the point estimate and its estimated standard error. For all methods except Rubin's, the standard error is multiplied by the normal percentile of 1.96; in Rubin's method, a t-multiplier is used with degrees of freedom determined from an approximate formula (1–3, 10). Confidence interval precision is measured by the ratio of upper and lower limits, which is the antilog of the width of the log odds-ratio confidence interval. For simulation efficiency, coverage of Rubin's method was computed by an indirect method which does not yield explicit limits, hence no ratio is given for Rubin's method.

Results

Table 1 presents results from a set of missing-at-random (MAR) simulations in which only the outcome influences the probability of missingness. As expected, all methods except ordinary missing indicator (OM) provide coverages reasonably close to the nominal 95 percent level; nevertheless, in the correlated situations, the coverage for regression imputation (RI) is subnominal with $p < 0.01$. When the covariates are correlated, the coverage of OM limits can be so poor that the ratio of confidence limits is not meaningful, except to show that the OM method tends to yield deceptively precise-looking results.

In this table and in all other situations examined, the differences between complete-subject (CS) and modified missing indicator (MM) was negligible, and the difference between the Rubin and Robins methods for analyzing multiple imputations (MI) was very small. Also, the efficiency of MI with only two imputations was not much different from that of complete-subject or modified indicator. Interestingly, the ratio of limits from missing-data maximum likelihood, theoretically the smallest achievable under the imputation model, was closely approached by the regression-imputation limits.

Table 2 involves a non-MAR missing-data mechanism in which the regressors but not the outcome influences missingness. The patterns it yields are nearly the same as in table 1, except that RI no longer yields subnominal coverage.

TABLE 1. MAR* simulations: X_1, X_2 bivariate normal, missingness independent of covariates but dependent on disease [$b = 0$, $d = \ln(4)$]

	Regressor correlation		
	−0.7	0	+0.7
Percent coverage			
CS*	94.6	95.4	95.2
OM*	55.5	95.6	91.8
MM*	94.7	95.4	94.8
RI*	93.5	94.7	93.1
MI*: Rubin†	95.0	96.7	95.4
Robins†	95.5	95.8	96.0
Mean odds-ratio estimate per unit regressor (true odds ratio = 2)			
CS	2.12	2.08	2.13
OM	1.54	2.01	2.32
MM	2.11	2.09	2.13
RI	2.02	2.01	2.10
MI: Rubin	2.11	2.06	2.03
Robins	2.10	2.05	2.02
Mean of upper/lower limit			
CS	2.55	2.09	2.80
OM	1.85	1.82	2.13
MM	2.55	2.09	2.80
RI	2.28	1.86	2.48
MI: Robins	2.63	1.96	2.76
ML*: missing data‡	2.26	1.81	2.39
all data	1.83	1.61	1.96

* MAR, missing-at-random; CS, complete-subject; OM, ordinary missing indicator; MM, modified missing indicator; RI, regression imputation; MI, multiple imputation; ML, maximum likelihood.
† Rubin, Rubin's method (1, 2); Robins, Robins' method (20).
‡ Robins and Gill (20).

Table 3 involves a non-MAR missing-data process in which both the regressors and the outcome influence missingness, and in which none of the methods yield valid confidence intervals. For these situations, the square roots of the mean-squared errors (root-mean-squared errors, or RMSE) replace the interval-width summaries. Although all the methods suffer from significant undercoverage, all but the ordinary indicator method retain over 90 percent coverage and less than 10 percent bias. Regression imputation is the least biased and has lowest RMSE, making it appear rather promising.

The latter promise is broken by the simulations in table 4, in which the regressors are lognormal but are misspecified as normal by the imputation methods. This misspecification leads to unacceptably large bias in naive regression imputation, a bias so large that for nonnegative correlations the regression imputation method does worse than the ordinary-indicator method. Yet, the same misspecification has relatively minor impact on the multiple-imputation method. Rubin's method now appears to have slightly better

TABLE 2. Non-MAR* simulations: X_1, X_2 bivariate normal, missingness depends on covariates but not disease [b = ln(1.2), d = 0]

	Regressor correlation		
	−0.7	0	+0.7
Percent coverage			
CS*	95.1	95.8	95.5
OM*	47.6	95.2	92.5
MM*	95.2	95.6	95.6
RI*	95.3	95.7	95.8
MI*: Rubin†	95.3	97.2	96.3
Robins†	96.8	96.3	97.0
Mean odds-ratio estimate per unit regressor (true odds ratio = 2)			
CS	2.10	2.07	2.10
OM	1.52	1.99	2.29
MM	2.10	2.06	2.09
RI	2.08	2.04	2.08
MI: Rubin	2.10	2.05	2.08
Robins	2.10	2.05	2.07
Mean of upper/lower limit			
CS	2.38	1.99	2.61
OM	1.77	1.75	2.04
MM	2.38	1.99	2.61
RI	2.28	1.86	2.47
MI: Robins	2.60	1.93	2.71
ML*: missing data‡	2.25	1.80	2.37
all data	1.83	1.61	1.95

* MAR, missing-at-random; CS, complete-subject; OM, ordinary missing indicator; MM, modified missing indicator; RI, regression imputation; MI, multiple imputation; ML, maximum likelihood.
† Rubin, Rubin's method (1, 2); Robins, Robins' method (20).
‡ Robins and Gill (20).

TABLE 3. Non-MAR* simulations: X_1, X_2 bivariate normal, missingness depends on covariates and disease [b = ln(1.2), d = ln(4)]

	Regressor correlation		
	−0.7	0	+0.7
Percent coverage			
CS*	92.2	91.0	93.0
OM*	41.2	85.9	95.8
MM*	92.2	90.8	93.0
RI*	92.2	93.4	90.3
MI*: Rubin†	91.6	91.5	94.1
Robins†	93.3	90.7	94.8
Mean odds-ratio estimate per unit regressor (true odds ratio = 2)			
CS	1.89	1.86	1.90
OM	1.43	1.81	2.04
MM	1.89	1.86	1.90
RI	1.93	1.98	2.11
MI: Rubin	1.89	1.88	1.94
Robins	1.89	1.88	1.94
Relative RMSE‡			
CS	1.65	1.68	1.63
OM	2.45	1.54	1.11
MM	1.65	1.69	1.64
RI	1.44	1.33	1.59
MI: Rubin	1.61	1.47	1.49
Robins	1.61	1.48	1.49

* MAR, missing-at-random; CS, complete-subject; OM, ordinary missing indicator; MM, modified missing indicator; RI, regression imputation; MI, multiple imputation.
† Rubin, Rubin's method (1, 2); Robins, Robins' method (20).
‡ RMSE (root mean-squared error) divided by RMSE of maximum likelihood estimator of b from all data.

coverage than Robins' method; this is apparently due to the inflation of the interval widths by the use of a t-multiplier (rather than a Z-multiplier) in Rubin's method. In terms of bias and RMSE, the complete-subject and modified-indicator methods also perform reasonably well.

Further simulations were conducted in which only one of the two covariates had missing values. The patterns observed above were unaltered, and so for brevity the results are omitted. As one might expect, the choice of missing-data method affected estimation of the complete covariate's coefficient in proportion to the correlation of that covariate with the partially missing covariate.

EXAMPLE: A RECORD-BASED STUDY OF ENDOMETRIAL CANCER

Medical-record studies conducted within a medical service plan or health maintenance organization may avoid certain problems such as selection bias (provided records are available on all members) and differential misclassification (provided records are completed without regard to the ultimate outcome of the member). Nonetheless, they often suffer from significant proportions of missing data due to incomplete records, and so missing-data methods can be of special importance for such studies.

Table 5 summarizes covariates and missing-data patterns from a record-based study of 318 endometrial-cancer cases and 599 controls in a prepaid health plan (24). Controls were matched to cases on age (3-year spans) and time in plan (5-year spans). Missing proportions ranged from zero (estrogen-use variables) to over 0.40 (age at menarche), with over half the subjects having at least one missing regressor. Conditional logistic regression models that included all the listed variables were fit to these data using four different methods: complete-subject, ordinary missing indicator, linear regression imputation, and Rubin's multiple imputation method. The latter employed 50 imputations based on a multivariate normal linear regression of incomplete regressors conditional on complete regressors, matching factors, and disease status.

TABLE 4. Non-MAR* simulations: X_1, X_2 misspecified as bivariate normal but actually lognormal, missingness depends on covariate and disease [b = ln(1.2), d = ln(4)]

	Regressor correlation		
	−0.4	0	+0.7
Percent coverage			
CS*	92.9	93.1	94.2
OM*	73.1	91.1	95.4
MM*	92.8	93.1	94.1
RI*	80.2	87.6	84.3
MI*: Rubin†	94.1	93.9	93.7
Robins†	92.6	91.9	92.1
Mean odds-ratio estimate per unit regressor (true odds ratio = 2)			
CS	1.94	1.88	2.03
OM	1.58	1.85	2.12
MM	1.93	1.87	2.02
RI	3.04	2.51	3.11
MI: Rubin	2.17	2.04	2.27
Robins	2.16	2.04	2.26
Relative RMSE			
CS	1.63	1.70	1.70
OM	1.65	1.44	1.08
MM	1.62	1.70	1.70
RI	2.65	2.22	2.49
MI: Rubin	1.61	1.56	1.85
Robins	1.60	1.56	1.84

* MAR, missing-at-random; CS, complete-subject; OM, ordinary missing indicator; MM, modified missing indicator; RI, regression imputation; MI, multiple imputation.
† Rubin, Rubin's method (1, 2); Robins, Robins' method (20).

TABLE 5. Missing data distribution in case-control study of endometrial cancer (318 cases, 599 controls from 37 pairs and 281 triplets matched on age and years in plan) (24)

Regressor	Percent missing	
	Cases	Controls
Unopposed estrogen use	0	0
Opposed estrogen use	0	0
Weight	1	2
Height	4	8
No. of abortions	3	5
No. of live births	2	5
Age at menarche	46	47
Age at first birth†	42	43
Age at menopause†	12	14
One or more of the above	53	59

* 11 variables giving number of prescriptions for every year before study up to 10, plus number of prescriptions over 10 years before study.
† Missing does *not* include nulliparous or premenopausal women, for whom these variables are set equal to age at study.

The large number of imputations yields standard errors close to those obtainable from the corresponding maximum-likelihood method (that is, maximum likelihood under a multivariate normal model for the incomplete regressors conditional on other variables).

Table 6 presents the estimated coefficients that were least and most sensitive to the estimation method, recent unopposed estrogen use (a weighted sum of the number of prescriptions in the 5 years before study), and age at menarche. For estrogen use, the choice of method makes no important difference, reflecting its relatively low correlation with incomplete regressors such as age at menarche. The loss of precision inherent in the complete-subject method is nevertheless apparent in its larger ratio of confidence limits.

For age at menarche, the complete-subject and missing-indicator methods produce results more compatible with positive effects than the results from the two imputation methods, and the imputation results are the more precise pair. Most striking is that the upper confidence limit from the missing-indicator method is 3.38, double that of the multiple-imputation method. Using lognormal rather than normal linear regressions had a negligible impact on the imputation results. Although the difference between the regression imputation and multiple imputation results are also small, the age-at-menarche confidence limits have a 22 percent higher ratio under multiple imputation than regression imputation. The apparently greater precision of the latter limits is illusory, however, because only multiple imputation takes account of the variability produced by the imputation process.

Although the results do not exhibit any dramatic sensitivity to choice of method, it would not have been possible to verify this fact until a range of methods

TABLE 6. Recent estrogen use (see text) and age at menarche: results from four methods

	Recent estrogen (1 unit)		Age at menarche (11 vs. 16 years)	
	Estimated odds ratio	95% CI*	Estimated odds ratio	95% CI
CS*†	1.60	1.10–2.34	1.42	0.91–2.22
OM*	1.96	1.44–2.66	1.54	0.70–3.38
RI*	1.82	1.35–2.46	1.07	0.78–1.46
MI*: Rubin‡	1.83	1.36–2.47	1.09	0.72–1.64
	p value		*p value*	
CS	0.014		0.43	
OM	0.00002		0.28	
RI	0.00007		0.68	
MI: Rubin	0.00007		0.69	

* CI, confidence interval; CS, complete-subject; OM, ordinary missing indicator; RI, regression imputation; MI, multiple imputation; df, degrees of freedom.
† 140 of 318 cases, 248 of 599 controls.
‡ Rubin, Rubin's method (1, 2); 50 imputations, approximate df >3,000 for all variables.

was applied. Given the theoretical and simulation superiority of multiple imputation, the multiple-imputation results were chosen for the final study report (24).

DISCUSSION

The simulations in tables 1–4 are of course much too limited to provide general positive guidelines. Nevertheless, they involve realistic parameter settings and thus serve as counter-examples to recommendations in favor of the ordinary indicator and regression imputation methods (4). They more tentatively suggest that if one is unable to carry out one of the more sophisticated approaches (such as multiple imputation, maximum likelihood, or weighted estimating equations), one may be best served by staying with the simplest method (complete-subject analysis) and avoiding intuitively appealing but potentially disastrous ad hoc approaches such as the ordinary indicator method or naive regression imputation. Because complete-subject analysis can also perform poorly (5), however, development of software packages for more sophisticated methods should be supported.

Multiple imputation based on a multivariate normal model for the incomplete regressors is thoroughly described in Rubin's textbook (1) and is not difficult to program. Rubin and Schenker (10) report that good confidence interval coverage can be obtained with just three imputations, and the simulations given here show that even two imputations can yield adequate coverage in some situations. Nonetheless, extra imputations are so easy to produce, once the necessary software is in hand, that we recommend use of 10 or more imputations in practice.

We have been informed that BMDP Statistical Software, Inc. (Los Angeles, California), is currently developing a multiple-imputation procedure for inclusion in their program series. In the interim, a GAUSS procedure (25) for generating multivariate normal imputations can be obtained by sending a self-addressed postpaid DOS disk mailer and 3.5-inch formatted diskette to the first author; this procedure requires GAUSS 3.2 or higher to run. Although multivariate normality is rare in epidemiology, imputation researchers to whom we have talked suggest that acceptable performance can be obtained by transforming all ordered regressors to approximate marginal normality for imputation, then transforming them back to their original scale or any other desired scale for the regression. Binary regressors are left as is, but their imputed values are rounded to the nearest of 0 and 1. An alternative to such an ad hoc imputation approach is to use more complex imputation models (1–3) or use maximum-likelihood or weighted estimating equation methods (8, 19). Unfortunately, software for such methods is not widely available.

Given that the joint distribution of regressors is rarely known and thus will be misspecified, results such as those in table 4 should encourage a reserved view of properties derived under the assumption of correct specification of the imputation model. Little and Rubin (2) express similar reservations about methods that require specification of the missing-data process. As in the situation they discuss, one can address the potential for specification bias by careful modeling and by examining sensitivity of results to changes in the models. Alternatively, one may employ imputation methods that depend less heavily on modeling, such as the matching method proposed by Heitjan and Little (11). For a recent discussion of related issues in multiple imputation, see Meng (26) and the accompanying comments.

ACKNOWLEDGMENTS

The authors thank Donald Rubin, Daniel Heitjan, Thomas Belin, Betz Halloran, Liu-Ping Zhao, and Girma Wolde-Tsadik for their helpful comments on the initial manuscript, and Harry Ziel for use of case-control data on endometrial cancer that were collected under NIH contract no. N01 CP 95680.

REFERENCES

1. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.
2. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley, 1989.
3. Little RJA. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227–37.
4. Miettinen OS. Theoretical epidemiology. New York: Wiley, 1985.
5. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol 1991;134:895–907.
6. Vach W, Blettner M. Erratum. Am J Epidemiol 1994;140:79.
7. Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. Technical report no. 94–2. Ames, IA: Department of Statistics, University of Iowa, 1994.
8. Robins JM, Zhao LP, Rotnitzky A. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 1994;89:846–6.
9. Chow WK. A look at various estimators in logistic models in the presence of missing values. In: Proceedings of the Business and Economics Section of the American Statistical Association. Alexandria, VA: American Statistical Association, 1979:417–20.
10. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991; 10:585–98.
11. Heitjan DF, Little RJA. Multiple imputation for the fatal accident reporting system. Appl Stat 1991;40:13–29.

12. Vach W, Schumacher M. Logistic regression with incompletely observed categorical covariates: a comparison of three approaches. Biometrika 1993;80:353–62.
13. Whittemore AS, Grosser S. Regression methods for data with incomplete covariates. In: Moolgavkar SH, Prentice RL. Modern statistical methods in chronic disease epidemiology. New York: Wiley, 1986:19–34.
14. Fay RE. Causal models for patterns on nonresponse. J Am Stat Assoc 1986;81:354–65.
15. Baker SG, Rosenberger WF, DerSimonian R. Closed-form estimates for missing counts in two-way contingency tables. Stat Med 1992; 11:643–57.
16. Carroll RJ, Gail MH, Lubin JH. Case-control studies with errors in covariates. Am Stat Assoc 1993;88:185–99.
17. Williamson GD, Haber M. Models for three-dimensional contingency tables with completely and partially cross-classified data. Biometrics 1994;50:194–203.
18. Wacholder S, Carroll RJ, Pee D, et al. The partial questionnaire design for case-control studies. Stat Med 1994;13: 623–34.
19. Vach W. Logistic regression with missing values in the covariates. New York: Springer, 1994.
20. Robins JM, Gill R. Non-response models for the analysis of ignorable missing data. Stat Med, in press.
21. Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. Stat Med 1991;10:739–47.
22. Fay RE. Comment. Stat Sci 1994;9:558–60.
23. Greenland S. A lower bound for the correlation of exponentiated bivariate normal pairs. Am Stat 1996;50 in press.
24. Finkle WD, Greenland S, Miettinen OS, et al. Endometrial cancer risk after discontinuing use of unopposed conjugated estrogens. Cancer Causes Control 1995;6:99–102.
25. GAUSS System 3.2. Maple Valley, WA: Aptech, 1994.
26. Meng X-L. Multiple-imputation inferences with uncongenial sources of input (with discussion). Stat Sci 1994;9:538–73.