doi:10.1111/j.1360-0443.2007.01946.x

Analysis of binary outcomes with missing data: missing = smoking, last observation carried forward, and a little multiple imputation

Donald Hedeker, Robin J. Mermelstein & Hakan Demirtas

University of Illinois at Chicago, Chicago, IL, USA

ABSTRACT

Aims Analysis of binary outcomes with missing data is a challenging problem in substance abuse studies. We consider this problem in a simple two-group design where interest centers on comparing the groups in terms of the binary outcome at a single timepoint. **Design** We describe how the deterministic assumptions of missing = smoking and last observation carried forward (LOCF) can be relaxed by allowing missingness to be related imperfectly to the binary outcome, either stratified on past values of the outcome or not. We also describe use of multiple imputation to take into account the uncertainty inherent in the imputed data. **Setting** Data were analyzed from a published smoking cessation study evaluating the effectiveness of adding group-based treatment adjuncts to an intervention comprised of a television program and self-help materials. **Participants** Participants were 489 smokers who registered for the television-based program and who indicated an interest in attending group-based meetings. **Measurements** The measurement of the smoking outcome was conducted via telephone interviews at post-intervention and at 24 months. **Findings and conclusions** The significance of the group effect did vary as a function of the assumed relationship between missingness and smoking. The 'conservative' missing = smoking assumption suggested a beneficial group effect on smoking cessation, which was confirmed via a sensitivity analysis only if an extreme odds ratio of 5 between missingness and smoking was assumed. This type of sensitivity analysis is crucial in determining the role that missing data play in arriving at a study's conclusions.

Keywords Dichotomous outcomes, imputation, LOCF, missing data, smoking.

Correspondence to: Donald Hedeker, Division of Epidemiology and Biostatistics (M/C 923), School of Public Health, University of Illinois at Chicago, 1603 West Taylor Street, Room 955, Chicago, IL 60612-4336, USA. E-mail: hedeker@uic.edu

Submitted 24 March 2006; initial review completed 24 July 2006; final version accepted 8 May 2007

INTRODUCTION

Binary outcomes are common in many research areas, notably in studies of drug use, alcohol and smoking. In these studies, it is the rule rather than the exception that the assessment of substance use (yes or no) will be unknown or missing for some subjects. How to handle these missing data has been a continuing source of controversy and debate [1,2]. Here, we consider this problem in a simple two-group design (e.g. control versus treatment), where the groups are compared in terms of a single binary outcome. In a longitudinal study, this might represent the outcome at the end of intervention or at the final follow-up. Also, we focus on smoking as the outcome, although the issues and methods described pertain to alcohol, drug use and other binary outcomes as well.

A common approach to dealing with missing smoking outcomes is to recode them as smoking. This is often called a 'conservative' assumption, in the sense that it is conservative to assume the worse of the two values of the outcome (i.e. smoking). Also, there is often a sense that some fraction of subjects might be missing because they are smoking, or in other words that missingness and smoking status are related.

Another relatively common imputation approach, that is used in studies with more than one time-point where interest centers on comparing the groups at a final time-point, is to recode the missing data on the final smoking outcome to the last available smoking

assessment. This procedure is termed 'last observation carried forward', or LOCF. Notice that under LOCF imputation, the actual timing of the 'final' smoking outcome can vary across individuals.

Deterministic imputations, like missing = smoking and LOCF, can be shown to yield severely biased results [1,3–5] for several reasons. First, if the missing data are related to group membership, then the comparison of groups on the outcome is confounded with the missing data. For example, if there are more missing data in the control group, then the 'conservative' missing = smoking assumption yields a test that favors the treatment group (unless it is true that smoking and missingness are perfectly related, as missing = smoking implies). Similarly, under LOCF, if time of dropout is related to group (i.e. control subjects drop out early, whereas treatment subjects drop out late or not at all), then group comparisons of the outcome are confounded with the amount and/or time of treatment. A further problem is that these deterministic imputations produce variance estimates that are artificially small, because the missing data are treated as a constant. This leads to biased estimates of association involving the recoded outcome variable, and measures of uncertainty (i.e. variances, standard errors) that are too small. Finally, the subsequent statistical analysis of the filled-in data (i.e. after the imputation has been performed) typically treats the imputed data as known, and so yields standard errors that are too small, *P*-values that are artificially low and rates of Type I error that are higher than nominal levels.

There has been a virtual explosion of development of statistical methods for dealing with missing data [6], and some excellent non-technical review articles have been published recently [4,7–11]. However, much of this development has not necessarily found its way into the substantive literature. One reason is that these methods may seem unduly complicated and are not always readily available in standard software, although this is certainly changing. Also, many of the developed approaches are for continuous outcomes, rather than binary outcomes, which are common in smoking, alcohol and drug studies.

In this paper, we aim at building upon the common naive imputation techniques, namely missing = smoking and LOCF, using a more statistically reasoned approach. Essentially, we will describe a multiple imputation approach that uses a variant of stochastic regression imputation [6]. However, because we are interested in relating missingness to smoking, and therefore do not assume that the data are 'missing at random' [12] or 'ignorable' [13], we will use missingness itself as a predictor of the imputed smoking status. Consequently, our approach will be the development of a non-ignorable imputation model that allows researchers to examine the sensitivity of the results to the relationship between missingness and smoking (for further description of these missing data terms, see [4] and [11]). Furthermore, because smoking status is binary, our approach is different to others in the addiction literature treating continuous outcomes [11,14]. Finally, because our intended audience is not statisticians, but researchers and data analysts, we describe and illustrate this approach in relatively non-technical terms. Our aim is to provide reasonable methods that can be implemented readily.

In terms of organization, we first describe relational versions of the missing = smoking and LOCF assumptions, respectively. In so doing, we illustrate why these deterministic imputations are logically problematic, and how they can be improved upon easily by positing them as non-perfect relationships. We further illustrate how these relationships can be cast within a logistic regression model, expressing the probability of smoking as a function of missingness and past behavior. We describe how this allows individual and sampling variability to be accounted for, and further allows development of a multiple imputation approach that accounts for the uncertainty inherent in the imputed data. We conclude with a discussion of pertinent issues.

THE RELATIONSHIP BETWEEN SMOKING AND MISSINGNESS

Suppose that there are two groups (treatment and control) and we are interested in whether smoking (yes or no) varies by group at a single time-point. A simple 2×2 table of group by smoking, with a Pearson's χ^2 test, can address this, but ordinarily it may be complicated by the fact that not everyone has a measurement on the smoking variable. In what follows, we will describe a way in which we can classify these missing individuals in the 2×2 table of group by smoking, based on an assumed relationship between missingness and smoking. For this, define Smoke as the binary dependent variable that is assessed at a single timepoint: either an individual is assessed to be smoking or not. Similarly, Miss is a binary indicator of whether the variable Smoke is assessed for a given individual: either an individual is assessed (i.e. it is known whether the individual is smoking or not) or the individual is not assessed (i.e. it is not known whether the individual is smoking or not). Then, we can create the simple 2×2 cross-tabulation of Miss by Smoke for the sample of n individuals presented in Table 1.

Here, n_{11} is the number of observed non-smoking individuals and n_{12} is the number of observed smoking individuals. The analogous frequencies in the second row n_{21} and n_{22} are unknown, because these are the numbers for individuals who are missing. What is known in the second row of Table 1 is $n_{2.}$, the total number of individuals who are missing. Note that, in Table 1 and

Table 1 Table of missingness (Miss) by smoking outcome (Smoke).

	Smoke		
Miss	*No*	*Yes*	*Total*
No	n_{11}	n_{12}	$n_{1.}$
Yes	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

subsequently, the 'dot' subscript indicates summation over the row or column.

While unknown, the numbers in the second row of Table 1 can be expressed as a simple function of the numbers in the first row and the odds ratio, denoted OR, which reflects the association between Miss and Smoke. Specifically, since the odds are expressed as $\text{OR} = (n_{22}/n_{21})/(n_{12}/n_{11})$, we can write

$$\frac{n_{22}}{n_{21}} = \text{OR}\frac{n_{12}}{n_{11}}. \quad (1)$$

The fraction on the left-hand side is the odds of smoking for missing individuals, while the similar fraction on the right-hand side is the odds of smoking for observed individuals. The missing = smoking assumption states that $n_{22} = n_{2.}$ and $n_{21} = 0$ (i.e. all missing observations are equal to smoking). This implies that the OR between Miss and Smoke goes to positive infinity, as the odds of smoking for missing individuals equals $n_{2.}/0$. Clearly, a more reasonable assumption would be that the OR, and the odds of smoking for missing individuals, is a finite value.

Because the ratio n_{12}/n_{11} is observed (i.e. the odds of smoking for non-missing individuals), equation 1 suggests that if a value for the OR is assumed, then one can calculate the number of missing individuals who are smoking and non-smoking. Specifically, the above equation implies that the number of missing individuals who are smoking equals:

$$n_{22} = n_{2.}\frac{\text{OR} \times odds_1}{1 + (\text{OR} \times odds_1)} = n_{2.}\pi, \quad (2)$$

where $odds_1$ denotes the odds of smoking for observed individuals (i.e. n_{12}/n_{11}) and π represents the indicated function of these odds and the OR. π is simply the probability of smoking for missing individuals under the assumed value of the OR and the observed odds of smoking for the non-missing subjects. Similarly, the number of missing individuals who are non-smoking equals $n_{2.}(1 - \pi)$.

Thus far, only missingness and smoking status have been considered. To bring group (i.e. control or treatment) into the picture, denote the number of missing individuals in the control group as $n_{2.c}$, and the number in the treatment group as $n_{2.t}$. Assuming that the relationship between missing and smoking is the same in the two groups, the number of missing control individuals who are smoking is calculated as $n_{22c} = n_{2.c}\pi$, and the number of analogous treatment individuals as $n_{22t} = n_{2.t}\pi$. Adding these calculated numbers to the observed frequencies, the usual χ^2 test for the association of smoking by group can be performed. This analysis is conditional on the assumed value for the OR, but sensitivity analysis can be performed by repeating this basic analysis under different values of the OR. In other words, the basic analysis of group by smoking is repeated under various assumed levels for the relationship between smoking and missing, and by so doing the robustness of conclusions concerning the group effect on smoking is assessed.

Example

Consider the data from the smoking cessation study described in Gruder *et al.* [15]. This study evaluated the effectiveness of adding group-based treatment adjuncts to a smoking cessation intervention comprised of a television program and self-help materials. Two types of group adjuncts were compared: one that included social support (buddy training) and relapse prevention training, and one that was a general group discussion of stopping smoking. The primary hypothesis of the study was that participants in the social support training condition would have higher abstinence rates than those in the general group discussion condition who, in turn, would do better at 24 months than participants who were in a condition that used only the television and self-help materials (no-contact control). Participants were smokers who registered for the television-based program (to receive the self-help materials) and who indicated an interest in attending group-based meetings. Although participants were randomized to condition, only 50% of the participants scheduled to attend a group session came to at least one meeting. Thus, for analysis, the original report considered four 'conditions': the two group treatments, a 'no show' group consisting of individuals who were randomized to a group treatment but never showed up for any group meetings, and the no-contact control condition. Measurement of the smoking outcome was conducted via telephone interviews, so individuals who did not attend group meetings were not necessarily missing in terms of the measurement of their smoking status.

Here, for simplicity and exposition, we combine the controls and no-shows as one group and the two active treatments as a second group. Hereafter, we refer to this first group as control and the second as treatment. This is akin to an 'as-treated' analysis which, as we will show, yields some interesting and varied results depending on

Table 2 Table of missingness (Miss) by Smoke, Gruder *et al*. [15] study.

	Smoke		
Miss	*No*	*Yes*	*Total*
No	78	294	372
Yes	n_{21}	n_{22}	117
Total	$n_{.1}$	$n_{.2}$	489

Table 3 Table of Group by Smoke under an odds ratio of 2 for Miss and Smoke.

	Smoke		
Group	*No*	*Yes*	*Total*
tx	38 + 3.9820	118 + 30.0180	190
Control	40 + 9.7207	176 + 73.2793	299
Total	91.7027	397.2973	489

the missing data assumptions. It should be noted that, in the face of treatment non-compliance, as-treated analyses do have problems, as do strictly 'as randomized' analyses; an excellent discussion of these issues, and some solutions that go beyond the scope of the present article can be found in [16]. Table 2 lists the frequencies for the cross-tabulation of missing by smoking status at the final study time-point (i.e. at 24 months).

For the observed individuals, the odds of smoking equal 294/78 = 3.77, or nearly 4–1. Clearly, the odds of smoking are quite high for observed individuals. The amount of missing data is rather large, accounting for approximately 24% of the total number of individuals. Thus, it would seem that the statistical analysis and resulting conclusions could be influenced by whatever is assumed for the missing data.

Analyzing only the available data (i.e. $n = 372$), the smoking rate in the control group is 176/216 = 81.48% and 118/156 = 75.64% in the treatment group. Conversely, the cessation rate is 18.52% in the control and 24.36% in the treatment group. There seems to be a slight advantage to the treatment group, but the simple Pearson's χ^2 test yields $\chi^2 = 1.86$, d.f. = 1, $P < 0.17$, which is not significant. Alternatively, if one assumes missing = smoking, then the smoking rates are 259/299 = 86.62% and 152/190 = 80.00% for the control and treatment groups, respectively, and we obtain $\chi^2 = 3.80$, d.f. = 1, $P < 0.051$, which is very nearly statistically significant. One might argue that a one-sided alternative is appropriate here (because we are only interested in whether treatment does better than control), yielding a P-value of 0.051/2 = 0.0255, and conclude that treatment is statistically superior to control under the 'conservative' missing = smoking assumption.

Why are the results from these two analyses different? Of the 117 missing individuals, 83 are from the control group and 34 from the treatment group, while the total number of control and treatment subjects equals 299 and 190, respectively. Thus, missingness is more common among the control group (83/299 = 27.8%) than the treatment group (34/190 = 17.9%), and so the 'conservative' missing = smoking assumption yields many more smokers in the control than the treatment group. Having more missing data in the control group is not unusual in smoking cessation studies (and in many other types of studies as well), where more attention and commitment is typically present for the treatment group, relative to the control group. Thus, 'conservatively' assuming these missing individuals are smoking does not necessarily yield a conservative test of the treatment effect. In fact, one might easily argue that it yields a rather liberal test of the treatment effect.

As an alternative to deterministically recoding missing to smoking, we can examine what happens under various assumed values of the OR for missing and smoking. For example, assuming that OR = 2 (i.e. the odds of smoking are double in the missing subjects than the non-missing subjects), we obtain:

$$\pi = \frac{2 \times 294/78}{1 + (2 \times 294/78)} = 0.8829, \quad (3)$$

implying an assumed smoking rate of 88.29% for missing individuals. The number of missing individuals who are smoking is calculated as $n_{22} = 0.8829 \times 117 = 103.2993$, and the number of non-smoking missing individuals is $n_{12} = 117 - 103.2993 = 13.7007$. These yield an odds of smoking of 7.54 for missing individuals (i.e. a doubling of the observed odds of smoking), implying that missing individuals have nearly an 8–1 odds of smoking under the OR = 2 assumption. Clearly, the probability of smoking is very high for missing individuals, although not equal to 1 as missing = smoking would imply. The number of missing smokers in the two groups can be calculated by multiplying this probability by the number of missing individuals in the group: $n_{22c} = 0.8829 \times 83 = 73.2793$ and $n_{22t} = 0.8829 \times 34 = 30.0180$. Using these calculated frequencies to augment the observed data yields the results presented in Table 3.

Here, smoking rates of 83.27% and 77.90% are obtained for control and treatment groups, respectively. The Pearson's χ^2 test yields $\chi^2 = 2.28$, d.f. = 1, $P < 0.131$, which is not significant. This calculation can be repeated under other assumed values of the OR to obtain a sense of the robustness of the results. In the present case, the two-tailed P-values diminish to 0.10, 0.09 and 0.08 as the OR

Table 4 Table of missingness (Miss) by Smoke stratified by prior smoking outcome (Smoke0).

	Smoke0 = non-smoking			*Smoke0 = smoking*		
	Smoke			*Smoke*		
Miss	*No*	*Yes*	*Total*	*No*	*Yes*	*Total*
No	$n_{111} = 42$	$n_{112} = 71$	$n_{11.} = 113$	$n_{211} = 36$	$n_{212} = 223$	$n_{21.} = 259$
Yes	n_{121}	n_{122}	$n_{12.} = 37$	n_{221}	n_{222}	$n_{22.} = 80$
Total	$n_{1.1}$	$n_{1.2}$	$n_{1..} = 150$	$n_{2.1}$	$n_{2.2}$	$n_{2..} = 339$

is increased to 3, 4 and 5, respectively. If one argues for a one-tailed test, and so divides these *P*-values in half, then the treatment group is seen to be significantly better than the control group only for OR between missing and smoking of 3 or higher (i.e. very strong association between missing and smoking).

PREVIOUS SMOKING INFORMATION

Thus far, we have based the probability of smoking as a function of missingness alone. Essentially, we have described a more relational version of the missing = smoking assumption. This permits missing individuals to be more likely to be smoking than observed individuals, and so rather than assuming that missingness and smoking is an absolute certainty, we can make it uncertain to various degrees.

Another simple imputation technique that is common in longitudinal studies is LOCF. Like missing = smoking, LOCF posits a perfect relationship in the sense that missing observations are related perfectly to previously observed values. Here, we use the idea underlying LOCF (i.e. that repeated observations are correlated), but again describe and develop it in a more relational manner. For this, consider the 2×2 cross-tabulation of missing by smoking status in Table 4, stratified by smoking status at a previous time-point.

In the current example the previous time-point is the post-intervention assessment, denoted as *t*0, and the smoking assessment at that time-point is designated as Smoke0. The time-point that the primary dependent variable to which Smoke corresponds is the final study assessment, 2 years after the post-intervention. Here, the order of the subscripts is table number, row, column, and the dot indicates summation over the indicated index. From Table 4, note that LOCF imputation would set all missing observations in the left-side table to non-smoking (i.e. $n_{121} = 37$ and $n_{222} = 0$) and all missing observations in the right-side table to smoking ($n_{221} = 0$ and $n_{222} = 80$). This would produce ORs going to zero and positive infinity in the two tables, respectively. Again, such deterministic imputation can be improved easily upon using the ideas of the previous section. Here, define:

$$\pi_i = \frac{\mathrm{OR}_i \times odds_{i1}}{1 + (\mathrm{OR}_i \times odds_{i1})} \quad (4)$$

where OR_i is the assumed OR for the *i*th table ($i = 1,2$), and $odds_{i1}$ are the observed odds of smoking for the *i*th table. π_i is the probability of smoking for the missing individuals under the assumed OR of the *i*th table. Note that above, the observed odds of smoking equal $71/42 = 1.6905$ and $223/36 = 6.1944$ depending on smoking status at *t*0. Clearly, the odds of smoking are much greater if previous smoking was observed, but the relationship is not perfect as LOCF would dictate.

LOCF would imply a zero OR for the left-hand table of Table 4 (i.e. the table for *t*0 non-smokers), but this would seem counterintuitive. Among those who are not smoking at *t*0, why would subjects who are missing at the final time-point have a lower chance of smoking than their observed counterparts? Instead, it would seem that while the observed odds of smoking should be different, depending on *t*0 smoking status, there does not seem to be any compelling reason as to why the OR for missing and smoking should be different for *t*0 smokers and non-smokers. In any case, the above equation permits the assumed ORs to vary, so one can examine this empirically.

In the present case, continuing with the assumption of OR = 2 for missing and smoking, assumed both for *t*0 smokers and non-smokers, we obtain:

$$\pi_1 = \frac{2 \times 71/42}{1 + (2 \times 71/42)} = 0.7717 \quad \text{and}$$

$$\pi_2 = \frac{2 \times 223/36}{1 + (2 \times 223/36)} = 0.9253$$

for the two tables, respectively. Thus, among those who are missing at the final time-point, subjects who were smoking at *t*0 have a very high assumed probability of smoking (0.925), while subjects who were not smoking at *t*0 have a lower smoking probability (0.772).

Using these values of π_i, and knowing the numbers of missing control and treatment subjects in each of these two tables, we can calculate the numbers of missing individuals who are smoking in the control and

Table 5 Table of Group by Smoke under an odds ratio of 2 for Miss and Smoke for both *t*0 smokers and non-smokers.

	Smoke		
Group	*No*	*Yes*	*Total*
tx	38 + 3.4245 + 1.4193	118 + 11.5755 + 17.5807	190
Control	40 + 5.0226 + 4.5567	176 + 16.9774 + 56.4433	299
Total	92.4231	396.5769	489

Table 6 Group by Smoke analyses under different missing data assumptions.

	Smoking frequencies (proportions) control	*Treatment*	χ^2	*P* <
Available data ($n = 372$)	176/216 (81.48)	118/156 (75.64)	1.87	0.17
Missing = smoking ($n = 489$)	259/299 (86.62)	152/190 (80.00)	3.80	0.051
Marginal OR = 1 ($n = 489$)	241.60/299 (80.80)	144.87/190 (76.25)	1.45	0.23
Marginal OR = 2 ($n = 489$)	249.28/299 (83.37)	148.02/190 (77.90)	2.28	0.13
Marginal OR = 5 ($n = 489$)	254.82/299 (85.22)	150.29/190 (79.10)	3.07	0.08
Stratified OR = 1 ($n = 489$)	242.34/299 (81.05)	143.78/190 (75.68)	2.02	0.16
Stratified OR = 2 ($n = 489$)	249.42/299 (83.42)	147.16/190 (77.45)	2.70	0.10
Stratified OR = 5 ($n = 489$)	254.76/299 (85.21)	149.82/190 (78.85)	3.28	0.07

OR = odds ratio for Miss and Smoke; marginal = marginal across all subjects; stratified = stratified by *t*0 smoking status.

treatment groups as: $n_{122c} = 0.7717 \times 22 = 16.9774$ and $n_{122t} = 0.7717 \times 15 = 11.5755$ for *t*0 non-smokers; and $n_{222c} = 0.9253 \times 61 = 56.4433$ and $n_{222t} = 0.9253 \times 19 = 17.5807$ for *t*0 smokers. Using these values to augment the observed data, we obtain the results listed in Table 5.

The smoking rates for the control and treatment groups now equal 83.42% and 77.45%, respectively, and the χ^2 test statistic is $\chi^2 = 2.70$, d.f. = 1, $P < 0.101$, which is nearly significant as a one-tailed test. Again, we can repeat this for other assumed values of the OR, additionally allowing the OR to vary across the two levels of Smoke0 if desired. Table 6 summarizes the results of this section for OR values of 1, 2 and 5.

An OR of 1 would imply that missing and smoking are independent, whereas an OR of 5 would imply a very strong, although imperfect, relationship. The results presented as 'marginal' do not take smoking status at *t*0 into account, whereas the 'stratified' results do, in the manner described. As can be seen from Table 6, the missing = smoking results are the most significant, although several others would be significant by a one-tailed test. This highlights the fact that the determination of whether the results are significant or not critically depends on the assumed OR for the association of missing and smoking. The missing = smoking assumption obscures this point, but the OR-dependent results bring this important point into focus.

MORE SOURCES OF VARIATION AND MULTIPLE IMPUTATION

While the imputation techniques described thus far represent a more informed approach to the problem of missing data, they are not ideal. One problem is that missing individuals are not allowed to have varying probabilities of smoking, other than permitting the probability of smoking to vary between the groups of *t*0 smokers and non-smokers. Thus, individual variation is being ignored. Another problem is that the sample proportion of smokers, either stratified or not by smoking status at *t*0, is treated as known, rather than as an estimate of a population parameter. Similarly, the OR, for the association of missing and smoking, is treated as known. Finally, the uncertainty inherent in the missing data is ignored, because the analysis does not distinguish between the observed and imputed data. Thus, the imputation strategy can be improved further by incorporating these sources of uncertainty into the process.

For this, we begin by framing our previous imputation approach in terms of a logistic regression model. Specifically, consider the following model for determining the number of individuals with Smoke = yes (#Smoking) and Smoke = no (#Non Smoking):

$$\log\left[\frac{\#\text{Smoking}}{\#\text{Non Smoking}}\right] = [\beta_0 + \beta_1 \text{Miss}][1 - \text{Smoke0}] + [\beta_2 + \beta_3 \text{Miss}]\text{Smoke0}. \quad (5)$$

Here, the covariate Miss is coded as 0 or 1 for observed or missing individuals, respectively. Similarly, previous smoking status at $t0$, denoted as Smoke0, is coded as 0 or 1 for non-smokers or smokers, respectively. Essentially, the previous imputations have been based on this model, using the observed data to estimate β_0 and β_2 (which reflect the odds of smoking based on the observed data for $t0$ non-smokers and smokers, respectively), and assuming values for β_1 and β_3 (which reflect the association of missing with smoking for $t0$ non-smokers and smokers, respectively). For example, in the last section we specified $\hat{\beta}_0 = 71/42$ and $\hat{\beta}_2 = 223/36$, and assumed $\beta_1 = \beta_3 = \log 2$ (i.e. the OR for missing and smoking is 2 for both $t0$ non-smokers and smokers, respectively), to calculate the log odds of smoking for missing subjects who were $t0$ non-smokers (i.e. $\hat{\beta}_0 + \beta_1$) and missing subjects who were $t1$ smokers (i.e. $\hat{\beta}_2 + \beta_3$). The numbers of missing smoking and non-smoking subjects can then be calculated directly from these log odds.

Individual variation

To bring individual variation into the picture, consider the following latent variable representation of this logistic regression model for subject i ($i = 1, \ldots, n$):

$$Y_i^* = [\beta_0 + \beta_1 \text{Miss}_i][1 - \text{Smoke0}_i] + [\beta_2 + \beta_3 \text{Miss}_i]\text{Smoke0}_i + \varepsilon_i \quad (6)$$

where ε_i represents an individual error term. Here, Y_i^* is a latent variable for subject i that is related to the observed binary smoking outcome y_i through the 'threshold concept' (see [17], pp. 40–44). Denoting the threshold or cut-point as γ, this concept states that if $Y^* > \gamma$ then $y = 1$, otherwise if $Y^* < \gamma$ then $y = 0$. This notion of an underlying latent variable y^* that manifests itself as a binary outcome y is not a critical assumption, although it can be used to motivate and describe the model. The logistic regression model is achieved by setting the threshold $\gamma = 0$ and specifying the error distribution as a standard logistic distribution, which has mean 0 and variance $\pi^2/3$ (see [17], p. 42).

Model 6 can be used to bring individual variation into the imputation process, allowing subjects with the same covariate values to have different probabilities of smoking. Specifically, one assigns a random draw for each missing subject from the distribution of ε_i (i.e. the standard logistic distribution), and then combines this random draw with a subject's calculated value of $x_i\beta$. This yields a value of Y^*, which then indicates $y = 1$ or $y = 0$ depending on whether or not Y^* is greater than zero (i.e. the assumed value of the threshold γ in the logistic regression model). This type of imputation is termed stochastic regression imputation [6] or imputing from a conditional distribution [4], albeit our use of the regressor Miss makes this imputation somewhat unique. Essentially, it allows us to perform a non-ignorable imputation in which missingness and the unobserved outcome are correlated (to the degree that β_1 and β_3 are non-zero). Performing this stochastic regression imputation once yields the results of Table 7.

Table 7 Table of Group by Smoke from stochastic regression imputation under an odds ratio of 2 for Miss and Smoke for both $t0$ smokers and non-smokers.

	Smoke		
Group	*No*	*Yes*	*Total*
tx	38 + 5	118 + 29	190
Control	40 + 10	176 + 73	299
Total	93	396	489

Here, as before, the imputed frequencies are indicated after the plus sign in each of the cells. Based on these data, the smoking rates are 83.28% and 77.37% for control and treatment groups, respectively, and the Pearson test value is $\chi^2 = 2.63$, d.f. = 1, $P < 0.105$. Note that these numbers are not very different from those based on the imputation of the previous section. However, as mentioned, the current imputation method is preferred statistically and logically because it allows for individual variation in the probability of smoking for those subjects with missing outcomes.

Sampling variation

Next, we need to take into account the fact that the values of the regression coefficients (i.e. the βs) are estimated or assumed, rather than known. Essentially, we need to allow for sampling variation in the imputation. For this, we can use standard results of logistic regression for the case of a single binary predictor [18]. Note that the logistic regression imputation model, given in equation 6 is of this type, stratified by $t0$ smoking status. In this case, the variances and covariance for the regression parameters associated with $t0$ non-smokers are given as:

$$V(\hat{\beta}_0) = (n_{111} + n_{112})/n_{111}n_{112}, \quad (7)$$

$$V(\hat{\beta}_1) = 1/n_{111} + 1/n_{112} + 1/n_{121} + 1/n_{122}, \quad (8)$$

$$C(\hat{\beta}_0, \hat{\beta}_1) = -(n_{111} + n_{112})/n_{111}n_{112}, \quad (9)$$

where the cell frequencies are specified as before. Notice that for the missing subjects, the frequencies n_{121} and n_{122} are obtained depending on the assumed level of the OR for missing and smoking.

The regression coefficients that are used in a given imputation need to incorporate these measures of

Table 8 Group by Smoke analyses under multiple imputation.

	Smoking frequencies (proportions) control	*Treatment*	χ^2	$P <$
Stratified OR = 1 ($n = 489$)	242.09/299 (80.97)	143.82/190 (75.70)	1.60	0.21
Stratified OR = 2 ($n = 489$)	248.87/299 (83.23)	146.95/190 (77.34)	2.28	0.13
Stratified OR = 5 ($n = 489$)	254.20/299 (85.02)	149.55/190 (78.71)	2.91	0.09

Averaged results based on 100 imputations. OR = odds ratio for Miss and Smoke; stratified = stratified by *t*0 smoking status.

uncertainty. For this, let $\hat{\beta}_{ns}$ denote the vector of estimated or assumed regression coefficients for *t*0 nonsmokers (i.e. $\hat{\beta}_0$ and β_1), and $V(\hat{\beta}_{ns})$ as its associated variance–covariance matrix (i.e. the matrix with elements in equations 7–9). For the regression coefficients used in a given imputation, which we denote as $\tilde{\beta}_{ns}$, we then take a random draw from a bivariate normal distribution with mean $\hat{\beta}_{ns}$ and variance–covariance $V(\hat{\beta}_{ns})$. Notice that, previously, we have set $\tilde{\beta}_{ns} = \hat{\beta}_{ns}$ and so assumed that the regression coefficients were known with certainty. By incorporating the uncertainty associated with these regression coefficients, we reflect the fact that they are unknown parameters whose values are either estimated or assumed.

This same process is used to obtain the regression coefficients for *t*0 smokers for a given imputation, denoted as $\tilde{\beta}_s$. Again, we draw from a bivariate normal with mean $\hat{\beta}_s$ (i.e. the vector of two regression coefficients for *t*0 smokers: $\hat{\beta}_2$ and β_3) and variance–covariance matrix $V(\hat{\beta}_s)$ (whose elements are obtained using equations 7–9, substituting frequencies from the table of *t*0 smokers, i.e. replacing n_{1ij} with n_{2ij}). These stochastic regression coefficients are then used in the model:

$$Y_i^* = [\tilde{\beta}_0 + \tilde{\beta}_1 \text{Miss}_i][1 - \text{Smoke0}_i] + [\tilde{\beta}_2 + \tilde{\beta}_3 \text{Miss}_i]\text{Smoke0}_i + \varepsilon_i \quad (10)$$

to yield imputed values for each missing observation that take into account both individual and sampling variation. As before, the continuous latent Y_i^* is dichotomized as y_i, based on the sign of the former.

Uncertainty attributable to missing data

Repeating this stochastic imputation many times, performing multiple imputation, is necessary in order to assess the variation that is attributable to imputation. This is critical in obtaining the *P*-value for the statistical test that combines the results for the multiply imputed data sets. The idea is that the *P*-value should reflect the uncertainty in the imputations of the missing observations. For instance, in our example, while the data for the observed subsample of 372 is certain, the data for the missing subsample of 117 is uncertain, and the *P*-value should reflect this uncertainty in the data. This point is essential and one that all the previous analyses glossed over.

Once the multiple imputations are performed, creating multiple versions of the data set, one can perform the statistical test of interest for each, and then combine the results over these repeated tests. The procedure for combining results from multiply imputed data sets is described in [19], and this procedure has been implemented recently in SAS PROC MIANALYZE. In the present case, we can perform a logistic regression of smoking status on group for each imputed data set, and then combine the results for testing the null hypothesis of the group effect on smoking. Table 8 presents the multiple imputation results considering assumed ORs of 1, 2 and 5 for smoking and missing.

The imputations were performed 100 times. As can be seen, the two-tailed *P*-values are now 0.21, 0.13 and 0.09 for OR of 1, 2 and 5. Note that these *P*-values are slightly larger than their analogous *P*-values in Table 6, where individual, sampling and imputation variations were not taken into account. Thus, taking into account stochastic imputation and variation attributable to it, the group effect on smoking is only significant under a one-tailed test if the assumed OR for missing and smoking exceeds approximately 5. Thus, under a one-tailed test, what was deemed statistically significant assuming missing = smoking is now recast as being significant only if one assumes a very strong relationship between missingness and smoking. This kind of sensitivity analysis is crucial in determining the role that missing data play in arriving at a study's conclusions.

DISCUSSION

In this paper, we have described a relatively simple approach for dealing with missing data in two-group studies with a binary outcome. Our focus has been to use the ideas behind missing = smoking and LOCF in a relational manner. In particular, it may be reasonable to assume that missing subjects are more likely to be smoking than observed subjects, and that smoking at a final time-point is related to smoking at a previous time-point. However, these relationships are not perfect ones.

Smoking status, in particular, may be more dynamic than is assumed frequently following a failed quit attempt, as relapsed smokers may quit again with the follow-up period of a study (e.g. [20]), or in more community-based population studies it is not uncommon to see abstinence rates increase over time (e.g. [21]). We have argued that it is important to examine results under a range of plausible values for the association of missing and smoking, stratified by past smoking behavior. By performing a sensitivity analysis one can determine the robustness of results to the assumed association of missing and smoking. Of course, if the conclusions vary across the range of this association, this does necessitate some kind of judgement call regarding what are plausible values for this association. In this regard, Schafer & Graham [4] note: 'one hopes that similar conclusions will follow from a variety of realistic alternative assumptions; when that does not happen, the sensitivity should be reported'. Thus, a major point of our argument is that blindly assuming missing = smoking provides a very unrealistic, if not absurd, solution to this dilemma. Uncertainty is perhaps difficult to deal with, but it is a fact of the matter when dealing with missing data.

Our approach advocates the use of multiple imputation, because individual, sampling and imputation variation can be accounted for. This is critical in obtaining a *P*-value for the test of treatment group by smoking that accurately reflects these real sources of variation. In our multiple imputation model we have posited that the primary smoking outcome is related to only two factors (i.e. missingness and past smoking behavior) and their potential interaction. Clearly, this is a simplistic approach, and a more sophisticated imputation model involving many more explanatory variables could be developed. For this, several articles have described more general multiple imputation models in psychological research [4,10,22], albeit primarily for continuous outcomes. While it is true that inclusion of additional variables to explain smoking in the imputation model can lead to improved results, missingness and past smoking behavior would seem to be very important predictors and certainly a good place to start in this endeavor. Furthermore, whereas additional covariates may or may not be present for different data sets, missingness is always available and past behavior is often available for a given data set. Also, in contrast to most imputation models, our model is a non-ignorable one, in that it allows missingness and the missed outcome to be correlated. This correlation is specified as the OR between missingness and smoking.

Multiple imputation is attractive for a number of reasons [19]. First, it works in conjunction with standard complete-data methods and software. Once the imputed data sets have been generated, the analyses can be carried out using procedures in virtually any statistical package. Secondly, one set of imputations may be used for a variety of analyses; there is often no need to re-impute when a new analysis is performed. Thirdly, the inferences—standard errors, *P*-values, etc.—obtained from multiple imputation are generally valid because they incorporate uncertainty due to missing data. Finally, multiple imputation can be highly efficient even if the number of imputations is relatively small, especially when between-imputation variance is not too large.

This paper has focused upon the simple analysis of smoking status at a single time-point, and the degree to which group differences in smoking status are present. Often, however, one is interested in examining smoking status over many time-points. Indeed, the study from which the data for this article were taken was such a longitudinal study [15]. For longitudinal smoking data, mixed-effects regression models offer a useful approach for analyzing incomplete data across time [23,24], albeit assuming the missing data are missing at random. Alternatively, a more sophisticated multiple imputation approach can augment the use of mixed-effects models for longitudinal data [3,25], although the methods are much more advanced than those presented in this article. In this paper, we have attempted to balance statistical sophistication with pragmatism to yield a sensible approach that can be applied readily, albeit for a relatively simple situation. For the interested reader, the data and computer syntax used in this article are available at http://www.uic.edu/~hedeker/long.html. In future work, we plan to examine how this approach can be extended for the analysis of longitudinal data.

Acknowledgements

Thanks are due to Siu Chi Wong for statistical analysis. This work was supported by National Institutes of Mental Health grant MH56146, National Cancer Institute grants CA80266 and 5PO1 CA98262, and by a grant from the Tobacco Etiology Research Network, funded by the Robert Wood Johnson Foundation.

References

1. Delucchi K. L. Methods for the analysis of binary outcome results in the presence of missing data. *J Consult Clin Psychol* 1994; **62**: 569–75.
2. Hall S. M., Delucchi K. L., Velicer W., Kahler C., Ranger-Moore J., Hedeker D. *et al.* Statistical analysis of randomized trials in tobacco treatment. *Nicotine Tob Res* 2001; **3**: 193–202.
3. Little R. J. A., Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics* 1996; **52**: 1324–33.
4. Schafer J. L., Graham J. W. Missing data: our view of the state of the art. *Psychol Methods* 2002; **7**: 147–77.
5. Cook R. J., Zeng L., Yi G. Y. Marginal analysis of incomplete

longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics* 2004; **60**: 820–8.
6. Little R. J. A., Rubin D. B. *Statistical Analysis with Missing Data*, 2nd edn. New York: Wiley; 2002.
7. Abraham W. T., Russell D. W. Missing data: a review of current methods and applications in epidemiological research. *Curr Opin Psychiatry* 2004; **17**: 315–21.
8. Collins L. M., Schafer J. L., Kam C.-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001; **6**: 330–51.
9. Gadbury G. L., Coffey C. S., Allison D. B. Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF. *Obestity Rev* 2003; **4**: 175–84.
10. Graham J. W., Cumsille P. E., Elek-Fisk E. Methods for handling missing data. In: Schinka J. A., Velicer W. F., editors (Weiner I. B. Editor-in-Chief). *Research Methods in Psychology, Volume 2 of Handbook of Psychology*. New York: Wiley; 2003, p. 87–114.
11. Yang X., Shoptaw S. Assessing missing data assumptions in longitudinal studies: an example using a smoking cessation trial. *Drug Alcohol Depend* 2005; **77**: 213–25.
12. Rubin D. B. Inference and missing data. *Biometrika* 1976; **63**: 581–92.
13. Laird N. M. Missing data in longitudinal studies. *Stat Med* 1988; **7**: 305–15.
14. Nich C., Carroll K. M. 'Intention to treat' meets 'missing data:' implications of alternative strategies for analyzing clinical trials data. *Drug Alcohol Depend* 2002; **68**: 121–30.
15. Gruder C. L., Mermelstein R. J., Kirkendol S., Hedeker D., Wong S. C., Schreckengost J. *et al.* Effects of social support and relapse prevention training as adjuncts to a televised smoking cessation intervention. *J Consult Clin Psychol* 1993; **61**: 113–20.
16. Little R. J. A., Yau L. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychol Methods* 1998; **3**: 147–59.
17. Long J. S. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications; 1997.
18. Agresti A. *Categorical Data Analysis*, 2nd edn. New York: Wiley; 2002.
19. Schafer J. L. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall; 1997.
20. Mermelstein R., Hedeker D., Wong S. C. Extended telephone counseling for smoking cessation: does content matter? *J Consult Clin Psychol* 2003; **71**: 565–74.
21. Turner L., Morera O., Johnson T., Crittenden K., Freels S., Parsons J. *et al.* Examining the effectiveness of a community-based self-help program to increase women's readiness for smoking cessation. *Am J Community Psychol* 2001; **29**: 465–91.
22. Sinharay S., Stern H. S., Russell D. The use of multiple imputation for the analysis of missing data. *Psychol Methods* 2001; **6**: 317–29.
23. Hedeker D., Mermelstein R. J. Application of random-effects regression models in relapse research. *Addiction* 1996; **91**: S211–29.
24. Hedeker D., Mermelstein R. J. Analysis of longitudinal substance use outcomes using ordinal random-effects regression models. *Addiction* 2000; **95**: S381–94.
25. Demirtas H. Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Stat Med* 2005; **24**: 2345–63.