Statistical Methods in Medical Research 2006; **15**: 213–234

Can one assess whether missing data are missing at random in medical studies?

Richard F Potthoff Duke Clinical Research Institute, Duke University Medical Center, Durham, NC, USA, **Gail E Tudor** Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA, **Karen S Pieper and Vic Hasselblad** Duke Clinical Research Institute, Duke University Medical Center, Durham, NC, USA

For handling missing data, newer methods such as those based on multiple imputation are generally more accurate than older ones and entail weaker assumptions. Yet most do assume that data are missing at random (MAR). The issue of assessing whether the MAR assumption holds to begin with has been largely ignored. In fact, no way to directly test MAR is available. We propose an alternate assumption, MAR+, that *can* be tested. MAR+ always implies MAR, so inability to reject MAR+ bodes well for MAR. In contrast, MAR implies MAR+ not universally, but under certain conditions that are often plausible; thus, rejection of MAR+ can raise suspicions about MAR. Our approach is applicable mainly to studies that are not longitudinal. We present five illustrative medical examples, in most of which it turns out that MAR+ fails. There are limits to the ability of sophisticated statistical methods to correct for missing data. Efforts to try to prevent missing data in the first place should therefore receive more attention in medical studies than they have heretofore attracted. If MAR+ is found to fail for a study whose data have already been gathered, extra caution may need to be exercised in the interpretation of the results.

1 Introduction

Missing data are a source of serious problems in statistical analyses of clinical trials and of other medical studies. In recent years, these problems have received increasing attention, with the development of improved methods for handling missing data. For example, the use of new techniques based on multiple imputation[1–3] has rightly been advocated as an improvement over single imputation and other simpler and earlier methods. Generally, multiple imputation based on ignorable nonresponse and a correct model is valid if the data are missing at random (MAR), whereas most simpler techniques require the data to be missing completely at random (MCAR) or require other assumptions that also are stronger than MAR.[4–9]

However, even the multiple-imputation methods based on ignorable nonresponse, as well as other sophisticated methods such as those based on maximum-likelihood estimation, are not valid if the MAR assumption fails. The use of these sophisticated methods[7,8] has been growing, with much written about their application (and with some work devoted also to sensitivity analysis and nonignorable modeling). But it appears that

Address for correspondence: Richard F Potthoff, Duke Clinical Research Institute, Duke University Medical Center, PO Box 17969, Durham, NC 27715, USA.
Gail E Tudor is currently with Husson College, Bangor, ME 04401, USA.

 10.1191/0962280206sm448oa

practitioners and theoreticians alike have paid little heed to the issue of trying to assess whether the MAR assumption holds in the first place. If one has no idea whether MAR holds (or to what extent it may be violated), then results of statistical analyses that assume MAR may be questionable. Of particular concern is a possible relationship between patients' healthiness and a tendency for data to be missing. If sicker patients are either more likely or less likely to have missing data, for example, then not only does the MCAR assumption fail, but also MAR becomes suspect.

Our motivation in writing this paper is to make people more aware of the pitfalls of missing data and the need to mention these matters in their writings. We see too many models run using only a subset of the original patients or using some form of imputation, but then no mention of the patients with missing data or the associated implications.

This paper addresses the issue of how one can try to detect whether MAR holds. If a statistical test yields evidence suggesting that MAR may fail, then one would hesitate to present results that rest on the MAR assumption; but if the test yields an absence of such evidence, then one can present the results with greater confidence. Although there is no way to directly test MAR itself, we propose an alternate assumption, to be called MAR+, that *can* be tested. (MAR+ is defined in Section 2.) MAR+ implies MAR, so a lack of evidence against MAR+ indicates no evidence contradicting MAR. Although MAR does not imply MAR+, we mathematically show (Propositions 1 and 2) that MAR and MAR+ are equivalent under certain conditions. That is, under these conditions, which are often plausible, the falsity of MAR+ implies the falsity of MAR.

This set of conditions is not satisfied, however, in a longitudinal study that has missing data in a monotone (ie, nested) response pattern. Thus, our results are not applicable to such longitudinal studies and are not intended for them. A paper by Little[10] concerned missing data in longitudinal studies and has a limited relation to our work.

In this paper, we make no attempt to assess the effects of MAR violations on statistical analyses that assume MAR, although such effort is important and has been undertaken in occasional earlier work (Section 6). We do, however, note the obvious fact that, regardless of whether or not MAR holds, missing data create no problems and no need for special techniques for any statistical analysis that involves only variables for which the data are complete. In particular, in a clinical trial in which patients are randomly assigned to two or more treatment groups that are to be compared with respect to an endpoint such as 30-day mortality, there will be no problem for an analysis that examines only the endpoint and uses no covariates, as long as the endpoint and the treatment group are both available for every patient. But if a secondary analysis of an endpoint uses covariates (such as sex, age, height, weight, pulse, blood pressure, smoking or hypertension), some of which are missing for some patients, then the question of whether MAR holds and the effects of MAR violations both become important.

For MCAR, unlike MAR, previous authors have already developed tests. The MCAR assumption means simply that missingness of data is independent of the values of any of the data. There seems to be a general belief that MCAR does not often hold in medical studies, in which case tests that assume MCAR would usually be invalid. Little[11] developed a test of MCAR for the case where all variables are continuous. Then Park and Davis[12] proposed a test of MCAR for the case where all variables are categorical. For some more recent work on MCAR tests, see Chen and Little[13] and Qu and Song.[14]

In the rest of the paper, we use a framework in which each patient has $(r + s)$ variables (dependent as well as independent) such that the first r are never missing and the remaining s are sometimes missing. Of the r never-missing variables, let the first r_1 be categorical and the last r_2 continuous, with $r = r_1 + r_2$. Because our methods do not work for $s = 1$, we assume $s > 1$. (For $s = 1$, though, MAR and MAR+ are the same.) Section 2 defines MAR+ and relates it to MAR, specifies for $s = 2$ and for general s the conditions under which MAR+ is equivalent to MAR, and proves this equivalence for $s = 2$ (for which the proof is simple) but relegates the proof for $s > 2$ to the appendix.

The main theme in Section 2 and the appendix is the mathematical proof that MAR and MAR+ are the same if certain (often tenable) conditions hold, whereas the rest of the paper deals with other topics: how to test MAR+, and empirical applications. We cover statistical tests of MAR+ in Section 3 for $s = 2$ and $r_2 = 0$, in Section 4 for general s and $r_2 = 0$ and in Section 5 for general s and $r_2 > 0$. At least one example based on real data appears in each of these sections; the tests in all but one example reject MAR+, though with differing degrees of conclusiveness. Concluding remarks discussing possible implications and limitations of our work are in Section 6.

2 MAR+ and its relation to MAR

In this section, we define MAR+ and present conditions under which it is equivalent to MAR. For each patient, there are r never-missing variables, which will be denoted by the vector $\boldsymbol{Z}(r \times 1)$, and s sometimes-missing variables. The special case of $s = 2$ will be dealt with to start, as it is the simplest case, but the case of general s will be covered after that.

For $s = 2$, let the two variables that are sometimes missing be called X and Y. Define a variable U that is 1 if X is missing and 0 if X is not missing. Define V to be 1 if Y is missing and 0 if Y is not missing. Then define the vector $\mathbf{m} = (u, v)$. Let $P(uv|x, y, \mathbf{z})$ denote the probability, conditional on x, y and $\mathbf{z}$, that $\mathbf{M} = (u, v)$. We will economize all succeeding notation by dropping the $\mathbf{z}$[eg, $P(uv|x, y, \mathbf{z})$ will appear as $P(uv|x, y)$], but it is to be understood that all probabilities will be construed as being conditional on $\mathbf{z}$.

If MAR holds together with the condition of *parameter distinctness*, then the missing-data mechanism is *ignorable*.[4–6,9] Definitions of MAR involve some subtle and complex elements. Consider the 2^s possible response patterns that indicate which of the s sometimes-missing variables are missing and which are not. For the development that follows, we define MAR to hold if the probability that a patient has any given pattern does not depend on the values of any of the variables that are missing for that pattern.

Thus, for MAR to be satisfied for $s = 2$ in our situation, $P(uv|x, y)$ can vary with a variable that is observed but not with the one that is missing. This means (cf. Example 1.13 of Little and Rubin[6, p. 18]) that $P(uv|x, y)$ must satisfy relations of the form

$$P(11|x, y) = f_{11}, \quad P(01|x, y) = f_{01}(x),$$
$$P(10|x, y) = f_{10}(y), \quad P(00|x, y) = f_{00}(x, y)$$

Note that these relations imply that

$$f_{11} + f_{01}(x) + f_{10}(y) + f_{00}(x, y) = 1 \quad \text{for all } (x, y)$$

(Remember that all probabilities are conditional on **z**, even though **z** is not shown.)

For $s = 2$, our alternative to MAR, which we call MAR+, requires that $P(uv|x, y)$ satisfy relations of the form

$$P(11|x, y) = k_{11}, \quad P(01|x, y) = k_{01}, \quad P(10|x, y) = k_{10}, \quad P(00|x, y) = k_{00}$$

with $k_{11} + k_{01} + k_{10} + k_{00} = 1$. In effect, the MAR+ condition is equivalent to stating that the data are MCAR (missing completely at random) for any fixed **z** or, in a manner of speaking, that the data are MCAR conditional on **z**. Of course, MAR+ is totally different from MCAR; MCAR implies MAR+, but (unless $r = 0$) MAR+ does not imply MCAR.

For $s > 2$, the definition of MAR+ is like that just given and thus remains simple. That is, the probability of any pattern of missingness is constant for any fixed **z** and never depends on any of the s sometimes-missing variables, whether observed or unobserved for that pattern.

Equation (3) of Vach and Blettner[15] is the same as our MAR+ condition. Those authors only mentioned the condition in a particular context, though.

Clearly, MAR+ implies MAR but MAR does not imply MAR+. In fact, it may appear that MAR+ is far more stringent than MAR. As it turns out, however, MAR and MAR+ are equivalent under conditions that may not be unreasonable at all in certain practical situations.

For $s = 2$, let $P(1 \cdot |x, y)$ and $P(\cdot 1|x, y)$ denote the respective probabilities that X and Y are missing, that is, that $U = 1$ and $V = 1$, conditional on x and y (as well as on **z**). Then, under MAR,

$$P(1 \cdot |x, y) = f_{11} + f_{10}(y) \quad \text{and} \quad P(\cdot 1|x, y) = f_{11} + f_{01}(x)$$

In addition, define $g_{**}(x, y)$ to be the probability, conditional on x and y and conditional also on one but not both of X and Y being missing, that X (rather than Y) is the variable that is missing. That is,

$$g_{**}(x, y) = \frac{P(10|x, y)}{P(10|x, y) + P(01|x, y)} = \frac{f_{10}(y)}{f_{10}(y) + f_{01}(x)}$$

under MAR.

Suppose that both X and Y are variables for which a higher value indicates, roughly speaking, a sicker patient. To simplify the exposition, here and later we speak of a 'sicker' patient, but with the understanding that this applies where one may expect sicker patients to have more (or at least no fewer) missing data. Data may be harder to gather for sicker patients and thus may be missing more often. In contrast, an opposite condition could also occur: less effort could be applied to obtain data for healthier patients if the data are deemed less important for their treatment. For situations where healthier patients may

be the ones with more missing data, our development still holds if one simply substitutes 'healthier' wherever 'sicker' appears. (More broadly, the development need not even be based on sicker or healthier patients. In some types of studies, it might be based on lower or higher socioeconomic status, for example.)

Now consider the following two conditions, both of which appear intuitively reasonable (in situations where sicker patients may be expected to have more, or at least no fewer, missing data).

Condition 1*. The probability that one variable is missing does not decrease as the patient registers sicker (ie, possesses a higher value) on the other variable. That is, $P(\cdot 1|x, y)$ is a nondecreasing function of x and $P(1 \cdot |x, y)$ is a nondecreasing function of y.

Condition 2*. Given that one but not both of X and Y are missing, the probability that X (rather than Y) is missing does not decrease as the patient registers sicker on X (with Y staying the same) and does not increase as the patient registers sicker on Y (with X staying the same). That is, $g_{**}(x, y)$ is a nondecreasing function of x for any fixed y and is a nonincreasing function of y for any fixed x. It will be assumed that neither $f_{10}(y)$ nor $f_{01}(x)$ is identically 0.

Observe that Condition 2* fails to hold in the case where X cannot be missing unless Y is also missing, because then $P(10|x, y) = f_{10}(y)$ is identically 0. In a longitudinal study in which X is measured at time 1 and Y at time 2, this case arises if a patient who misses the X measurement is automatically dropped from the study (indicating a monotone response pattern), but it does not arise if the patient can rejoin the study after missing the X measurement. The presence of a patient with $U = 1$ and $V = 0$ shows that $f_{10}(y)$ is not identically 0, but note that, because $f_{10}(y)$ is a probability, the absence of such a patient in a given data set does not suffice, in and of itself, to show the opposite. For general s, remarks similar to the foregoing apply for Condition 2 below.

Proposition 1. *For $s = 2$, if Conditions 1* and 2* both hold, then MAR implies MAR+ and is thus equivalent to MAR+.*

Proof of Proposition 1. First, Condition 1* obviously implies, under MAR, that $f_{01}(x)$ is nondecreasing in x and $f_{10}(y)$ is nondecreasing in y. Second, from Condition 2* [including the stipulation that neither $f_{10}(y)$ nor $f_{01}(x)$ is identically 0], it follows that $f_{01}(x)$ is nonincreasing in x and $f_{10}(y)$ is nonincreasing in y. Thus, taken together, these two pairs of results establish that $f_{01}(x)$ and $f_{10}(y)$ are both constant. Moreover, if $f_{01}(x)$ and $f_{10}(y)$ are constant, then $f_{00}(x, y)$ also has to be constant, because the four $P(uv|x, y)$ probabilities must add to 1 for all (x, y). The conclusion is, therefore, that MAR+ is satisfied.

Next, the above development can be extended to general s by generalizing Conditions 1* and 2* as follows.

Condition 1. If A is any proper, non-empty subset of the s sometimes-missing variables, then the probability that the variables in A are all missing does not decrease

as the patient registers sicker on any sometimes-missing variable not in A (with any remaining sometimes-missing variables not in A staying the same, at any fixed values).

Condition 2. Let X' and X'' be any two of the s sometimes-missing variables. Let $\mathbf{x}_{s-2}[(s-2)\times 1]$ be the values of the remaining $(s-2)$ sometimes-missing variables, and let $\mathbf{m}_{s-2}[(s-2)\times 1]$ denote any of the 2^{s-2} possible vectors of 1's and 0's that represent patterns in which these $(s-2)$ variables can be missing or present (1's stand for missing, 0's for present). Then the condition is:

> Given that one but not both of X' and X'' are missing, and conditional also on $\mathbf{m}_{s-2}$ and on $\mathbf{x}_{s-2}$, the probability that X' (rather than X'') is missing does not decrease as the patient registers sicker on X' (with X'' staying the same) and does not increase as the patient registers sicker on X'' (with X' staying the same). That is, this conditional probability (g) is a nondecreasing function of X' for any fixed X'' and is a nonincreasing function of X'' for any fixed X'. As in Condition 2*, it is assumed that neither of the probabilities used in obtaining g is identically 0.

As with the case of $s = 2$, all probabilities for the case of general s are to be interpreted as being conditional on $\mathbf{z}(r\times 1)$, the values of the r never-missing variables.

Proposition 2. *For general s, if Conditions 1 and 2 both hold, then MAR implies MAR+ and is thus equivalent to MAR+.*

In order to avoid complicated notation, the proof of Proposition 2 will be given only for $s = 3$. From the proof for $s = 3$, however, the extension to a proof for general s should be readily apparent. The proof for $s = 3$ is in the Appendix.

Because of Proposition 2, a statistical test result that rejects MAR+ will cast doubt on the use of missing-data techniques that assume MAR, if Conditions 1 and 2 are reasonable. In contrast, a result that fails to reject MAR+ will enhance one's confidence in any technique that assumes MAR, regardless of whether Conditions 1 and 2 hold.

3 Statistical tests of MAR+ for $s = 2$ and $r_2 = 0$

Although MAR cannot be tested statistically (except possibly if extra assumptions are made), MAR+ *can* be tested. To start our discussion of tests of MAR+, we examine the simplest case. This is the case where there are only $s = 2$ sometimes-missing variables and where all r of the never-missing variables are categorical (ie, $r_2 = 0$ of the never-missing variables are continuous and $r_1 = r$ are categorical).

Let X and Y be the two sometimes-missing variables, with a focus on those patients for whom X is not missing. Then MAR+ implies (among other things) that, conditional on the values of the r never-missing variables in $\mathbf{Z}$ and conditional also on X not being missing, the probability of Y being missing is a constant and, in particular, does not depend on the value of X. To put it another way, for given $\mathbf{z}$ and for X not missing, the value of X is independent of whether Y is missing or not, under MAR+. Thus one can test MAR+ if one tests, for fixed $\mathbf{z}$, whether the distribution of observed (nonmissing) X is the same for patients with Y missing as for those with Y present.

Suppose i) that the number of cells formed by combinations of values of the never-missing variables in **Z** (all r of which are categorical) is not excessive in relation to the number of observations (patients). Suppose also ii) that X is dichotomous, is otherwise ordered and categorical or is continuous.

Then a reasonable way to test MAR+ for the ordered pair (X, Y) of sometimes-missing variables is to calculate the Wilcoxon two-sample rank-sum statistic for each cell (X for Y missing versus X for Y present) and obtain a weighted sum of the resulting Wilcoxon statistics across cells. One then subtracts the null expectation of the weighted sum, divides the difference by the square root of the null variance and refers the resulting quotient to the normal distribution to find the P-value. If the weight for cell (or stratum) h is $1/(n_h + 1)$, where n_h is the number of patients in the cell, then the weights are the same as those of van Elteren,[16] also given in Equation (3.24) of Lehmann.[17]

Note that the development in the three preceding paragraphs still holds if the roles of X and Y are reversed. Thus there are two tests. One of them assesses whether observed X has the same distribution whether Y is missing or present, and the other assesses whether observed Y is distributed the same whether X is missing or present. Through the use of Bonferroni's inequality, the two tests can be combined into a single overall test by doubling the smaller of the two P-values.

Example 1. A set of data presented in Table 1 of Fuchs[18] has some missing values and, in earlier analyses by other authors, evoked considerable controversy stemming in part from the handling of the missing data. The data set came from a randomized

Table 1 Data for Example 1

Cell	Never-missing variables				Number of persons with mental status (Y)			
					Present ($V = 0$) and with physical status		Missing ($V = 1$) and with physical status	
	Vital status (deceased or survived)	Sex (male or female)	Age [younger (<75) or older]	Group (experi-mental or control)	Good ($X = 0$)	Poor ($X = 1$)	Good ($X = 0$)	Poor ($X = 1$)
1	D	M	Y	E	1	0	1	2
2	D	M	Y	C	1	2	0	0
3	D	M	O	E	6	8	0	5
4	D	M	O	C	8	4	0	3
5	D	F	Y	E	3	0	0	1
6	D	F	Y	C	1	0	0	1
7	D	F	O	E	1	3	1	2
8	D	F	O	C	2	3	2	0
9	S	M	Y	E	5	4	1	1
10	S	M	Y	C	12	1	3	1
11	S	M	O	E	8	1	2	0
12	S	M	O	C	8	2	1	3
13	S	F	Y	E	3	0	1	0
14	S	F	Y	C	5	1	1	0
15	S	F	O	E	2	1	0	0
16	S	F	O	C	4	1	0	1
All					70	31	13	20

controlled study of elderly people who had trouble caring for themselves and were referred to community agencies for protective help. The subjects were assigned either to an experimental group that received enriched social casework services or to a control group. At issue, and the subject of 'heated debate,'[18, p. 277] was whether or not the special services were 'detrimental'[18, p. 274] with respect to mortality. On the basis of special procedures that assumed MAR, Fuchs[18] found no negative effect of the services but also cautioned (p. 274) that "we do not claim to settle the debate, since we have no information to ensure ourselves that in this study the missing data are indeed 'missing at random' as is assumed in our analyses."

The data have $s = 2$ sometimes-missing variables (mental status and physical status, each classified as good or poor) and $r = 4$ never-missing variables (deceased or survived, the sole outcome variable among the six; male or female; younger or older, with only a binary age split being available from the published data; and experimental or control group). All six variables are dichotomous. Thus $r_1 = 4, r_2 = 0$ and the number of possible cells with different $\mathbf{z}$ is 2^4, or 16. There are data for 164 persons, of whom 101 have neither mental nor physical status missing, 33 have just mental status missing, one is missing only physical status and 29 are lacking both. Our Table 1 gives a rearrangement of the data from Fuchs for the 134 participants for whom physical status is not missing. For each of the 16 cells, there is one row, which shows, separately for mental status (Y) present and missing, the number of participants with good and poor physical status (X).

Each row (cell or stratum) can yield a Wilcoxon statistic. Cells 2, 13 and 15 are excluded, however, either because all participants have the same physical status or because all persons are alike with respect to whether mental status is present or missing. Cell 6, in contrast, is retained even though it has only two persons, because these two are opposite with respect to both physical status and the presence or absence of mental status. Note that, where two or more values of X are tied within a cell, the *midrank*,[17] or mean of the ranks associated with the tied values, is used in the calculations.

For cell 12 (chosen for illustration), the Wilcoxon statistic is taken as the sum of midranks of physical status for the four persons with mental status missing. It is thus equal to $(1 \times 5) + (3 \times 12)$, or 41, if one assigns a larger rank for poor (versus good) physical status. A weight of $1/(14 + 1)$ is applied, yielding 41/15. The contribution of cell 12 to the null expectation of the weighted sum of the Wilcoxon statistics is $(4 \times 15)/2$ divided by 15, or 2. On the basis of Formula (1.35) of Lehmann,[17] which takes account of ties, cell 12 contributes 450/13 divided by 15^2, or 2/13, to the null variance of the weighted sum.

Upon making similar calculations for the other 12 contributing cells and then summing across all 13 cells, one finds that the weighted sum is 19.394, its null expectation is 16.5 and its null variance is 1.0026. Thus one refers +2.894 divided by 1.0013, or +2.890, to the normal distribution to obtain a two-tailed P-value of 0.00385.

The next step is to trade X and Y and try to assess whether (conditional on $\mathbf{z}$) the distribution of observed mental status is the same regardless of whether physical status is missing or present. As indicated earlier, however, the data set[18] has only one person with physical status missing and mental status not missing. That person is in the (S, M, O, E) cell and his mental status is poor. In the same cell are nine more persons with mental status not missing (but with physical status also nonmissing). Mental status is good for

seven of them and poor for the remaining two. Thus, the cell has values (analogous to those of Table 1) of (7, 2, 0, 1). The *P*-value for *X* and *Y* traded has to be calculated only from this single cell. It far exceeds 0.00385, regardless of whether it is calculated exactly based on the hypergeometric distribution (which gives $P = 0.3$ one-tailed) or otherwise.

Thus, when one applies the Bonferroni inequality, the *P*-value for the final, overall test is twice 0.00385, or 0.0077. We conclude that, for Example 1, the assumption of MAR is highly suspect (if one deems Conditions 1* and 2* to be reasonable) and one should adopt a wary stance toward any results based on missing-data techniques that assume MAR. Fuchs's warning, quoted earlier, does not seem overly cautious.

The *P*-value of 0.0077 is two-tailed and would be halved if one were to apply a one-tailed test by invoking a prior supposition that missingness could be positively but not negatively associated with poorer medical condition. In our later examples as well as this one, however, we will adhere to two-tailed tests.

4 Statistical tests of MAR+ for $s \geq 2$ and $r_2 = 0$

We now generalize the development of the preceding section and consider the case where there can be any number (≥ 2) of sometimes-missing variables but where all r of the never-missing variables are still categorical. We invoke the same suppositions, i) and ii), as in Section 3. To begin, define the variable D to be the number of the s sometimes-missing variables that are missing for a given patient, so that D is an integer ≥ 0 but $< s$. (A patient could also have $D = s$ but would then be excluded from all statistical tests.) Let X be any one of the s sometimes-missing variables. We employ an argument like that in the second paragraph of Section 3. Conditional on $\mathbf{z}$ and conditional on X not being missing, the distribution of D does not depend on the value of X under MAR+. Thus, one can test MAR+ if, for fixed $\mathbf{z}$, one tests for independence between X and D, among those patients for whom X is not missing.

It is not our purpose to do a detailed assessment to try to find the best test of independence between X and D. Rather, we suggest one test, the Mantel–Haenszel-type statistic $Q_{\mathrm{MH}(3)}$ defined by Equation (A5) of Landis *et al.*,[19] that appears reasonable and can readily be used in all of our examples. Its null distribution is approximately chi-square with one degree of freedom. To calculate $Q_{\mathrm{MH}(3)}$, one has to choose[19] two sets of score or weight vectors, $\mathbf{c}_h$ and $\mathbf{a}_h$, that apply, respectively, to the levels or values of X and D. Two options that are frequently favored are based on ridit-type[20, p. 210] and modified-ridit-type[21, p. 65] scores, which are the ranks (midranks in case of ties) divided by n_h and $(n_h + 1)$, respectively. Our choice will be a hybrid of these two: for both $\mathbf{c}_h$ and $\mathbf{a}_h$ we will use the ranks (or midranks) divided by the square root of $n_h(n_h + 1)$. (Equivalently, the same value of $Q_{\mathrm{MH}(3)}$ will result if one uses ridit-type scores for $\mathbf{c}_h$ and modified-ridit-type scores for $\mathbf{a}_h$, or vice versa.)

For $s = 2$, the $Q_{\mathrm{MH}(3)}$ test as just described is equivalent to the test based on the weighted sum of Wilcoxon statistics that was defined in Section 3. This equivalence, which holds regardless of the extent of any ties in the X variable, can be shown with some rather lengthy algebra. Our $Q_{\mathrm{MH}(3)}$ test is thus a generalization, for $s \geq 2$, of the Wilcoxon-type test that was presented for $s = 2$.

One can apply our $Q_{MH(3)}$ test to each of the s sometimes-missing variables separately, by taking each of these s variables to be the X variable. One thus calculates s different P-values, the smallest of which can be multiplied by s to obtain a single P-value that, by virtue of the Bonferroni inequality, provides one overall test of MAR+. Although there are alternatives to the Bonferroni inequality that may be somewhat preferable, we apply it in our examples to make our presentation simpler.

We point out, but will not consider closely, some alternatives to our $Q_{MH(3)}$ test. One might generalize the Wilcoxon-based test along the lines of the Jonckheere–Terpstra test. One could use $Q_{MH(3)}$ but with a different choice for $\mathbf{c}_h$ and $\mathbf{a}_h$.

A third option is a weighted sum of the X-versus-D Spearman rank correlation coefficients from the different cells. It follows from Equation (7.26) of Lehmann[17] that the null variance of a Spearman statistic is $1/(n_h - 1)$, regardless of whether or not there are ties. Thus, it is natural to use weights of $(n_h - 1)$. The weighted sum divided by the square root of its null variance of $\Sigma_h(n_h - 1)$ is referred to the normal distribution to obtain the P-value. One can show that, for any value of s, this Spearman-based test is identical to our $Q_{MH(3)}$ test in the (unusual) case where no cell has any ties for either X or D.

We have not yet dealt with the situation where the variable X is categorical but is not ordered and not suitable to be treated as ordered. Even though this situation is the principal one that does not conform to supposition ii) of Section 3, in medical applications there may be few variables that are categorical and unordered, and none of our examples has a variable of this type. If such a variable (eg, country) is encountered, though, one approach would be to test for independence of X and D through the generalized Kruskal–Wallis statistic $Q_{MH(2)}$ defined by Equation (A4) of Landis *et al*.[19] This statistic requires a score vector $\mathbf{a}_h$ to apply to D (but no vector $\mathbf{c}_h$ to apply to X). For $\mathbf{a}_h$, one might use the ranks (or midranks) divided by the square root of $n_h(n_h + 1)$.

Example 2. This example uses the data in Table 1 of Little and Schluchter,[22] which came from a study of psychological risk in 69 families with two children each. Little and Rubin[6, p. 228 ff.] and Schafer[9, pp. 359–367] also analysed these data. These are $s = 6$ sometimes-missing variables (listed in Table 2) but just one never-missing variable, a parental risk group classified according to three ordered risk categories based largely on the presence and extent of parental psychiatric illness.

Table 2 Results for the $s = 6$ sometimes-missing variables of Example 2

Variable	Type of variable	Percentage of data missing (%)	$Q_{MH(3)}$ P-value	Sign of association
First child's score on				
Reading	Continuous	30	0.80	−
Verbal comprehension	Continuous	43	0.31	−
Symptom severity	Dichotomous	41	0.66	+
Second child's score on				
Reading	Continuous	23	0.26	+
Verbal comprehension	Continuous	25	0.72	−
Symptom severity	Dichotomous	41	0.24	+

$Q_{MH(3)}$ for X versus D was calculated as described earlier, using each sometimes-missing variable as X. Table 2 shows the six resulting P-values. The smallest one, 0.24, lacks significance even before multiplication by s to apply the Bonferroni inequality. Thus, there is no reason to reject the null hypothesis that MAR+ holds nor to suspect that MAR is violated. Of course, the small sample size necessarily limits the power of the test.

One should take note of the sign of the term in the numerator of $Q_{MH(3)}$ before this term is squared, so that one knows whether X and D are positively or negatively associated. The last column of Table 2 shows this sign. A plus (minus) sign in the table indicates a positive (negative) association between missing data, that is, high D, and a high value of X. Thus, for example, families whose children had more severe symptoms were more likely to have missing values.

Example 3. Data for this example come from a subset of 2972 patients taken from the multicenter Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO-I) randomized clinical trial of 41021 patients with acute myocardial infarction. The subset[23] consists of those patients with cardiogenic shock, present either at the outset (11%) or only after admission to the hospital (89%). Our data set has $s = 21$ sometimes-missing variables (shown in Table 3) and $r = 3$ never-missing variables, all of which are categorical ($r_1 = 3, r_2 = 0$).

Note that the 21 sometimes-missing variables include three that are categorical with more than two groups. We chose to treat all three as ordered rather than unordered categorical variables, although one might debate these choices.

The never-missing variables are drug treatment, which has four unordered categories; geographic area, either United States or other; and sex. Thus, the number of cells with different $\mathbf{z}$ is $4 \times 2 \times 2$, or 16.

As in Example 2, $Q_{MH(3)}$ for X versus D was calculated using each sometimes-missing variable as X. The 21 resulting P-values, as well as the signs of the association between X and D, appear in Table 3, both under a heading '$n = 2972$.'

The P-values and signs that appear under a heading '$n = 2968$' in Table 3 refer to a second run that we made. The 30-day-mortality variable (which could be treated as a response variable to be predicted from the remaining variables) was missing for only four patients. To explore what would happen, we removed these four patients for the second run, leaving 2968 patients. The dropping of the four patients, however, has less impact than the change of 30-day mortality from a sometimes-missing to a never-missing variable. Because there are now four never-missing and 20 sometimes-missing variables rather than three and 21, the number of cells with different $\mathbf{z}$ becomes 32 instead of 16.

The P-values are below 0.00001 for systolic blood pressure, diastolic blood pressure and Killip class in both runs and for 30-day mortality in the first run (the only run for which it has a P-value). A greater amount of missingness is associated with lower systolic and diastolic blood pressure and with higher Killip class and 30-day mortality. The P-value for weight is under 0.001 in both runs; lower weight is associated with greater missingness. For either run, after applying the Bonferroni inequality by multiplying the smallest P-value by 21 or 20, one easily rejects the null hypothesis that MAR+ holds.

Table 3 Results for the $s = 21$ sometimes-missing variables of Example 3

Variable	Type of variable	Percentage of data missing (%)	$Q_{MH(3)}$ *P*-value		Sign of association	
			$n = 2972$	$n = 2968$	$n = 2972$	$n = 2968$
Weight	Continuous	6	0.00005	0.00030	−	−
Height	Continuous	19	0.66	0.87	−	−
Systolic blood pressure	Continuous	3	<0.00001	<0.00001	−	−
Diastolic blood pressure	Continuous	8	<0.00001	<0.00001	−	−
Heart rate	Continuous	4	0.15	0.89	+	+
Time to treatment	Continuous	6	0.67	0.09	−	−
Age	Continuous	0.2	0.00012	0.79	+	+
Smoking history (1, current; 2, former; 3, never)	Ordered categorical	5	0.91	0.02	−	−
Race (1, white; 3, black; 2, other)	Ordered categorical	3	0.97	0.77	+	−
Killip class (I, II, III, IV, coded as 1, 2, 3, 4)	Ordered categorical	1	<0.00001	<0.00001	+	+
Infarction location (1, anterior; 0, all other)	Dichotomous	0.3	0.84	0.03	−	−
Hypertension[a]	Dichotomous	1	0.76	0.47	−	−
Diabetes[a]	Dichotomous	1	0.0015	0.11	+	+
Previous myocardial infarction[a]	Dichotomous	1	0.06	0.75	+	+
Previous angina[a]	Dichotomous	2	0.64	0.47	+	−
Prior cerebrovascular disease[a]	Dichotomous	1	0.09	0.18	+	+
Prior bypass surgery[a]	Dichotomous	1	0.89	0.87	+	−
Prior angioplasty[a]	Dichotomous	1	0.83	0.92	+	−
Hypercholesterolemia[a]	Dichotomous	7	0.04	0.22	−	−
Family history of coronary heart disease[a]	Dichotomous	11	0.38	0.96	−	−
30-day mortality (0, survived; 1, deceased)	Dichotomous	0.1	<0.00001	none	+	none

[a]Coded 1, condition present; 0, condition absent.

The most striking difference between the two runs is that the *P*-value for age is slightly above 0.0001 in the first run but far from significant in the second run. In addition, the *P*-value for diabetes is 0.0015 in the first run but 0.11 in the second, and for hypercholesterolemia it is below 0.05 in the first run but not in the second. In contrast, both smoking history and infarction location have *P*-values that are under 0.05 in the second run but nowhere close to significance in the first run. Except for four variables whose *P*-values exceed 0.45 in both runs, the signs of the association between each variable and missingness are the same in the two runs.

The differences noted between the two runs may not be surprising in view of the fact that whether or not MAR+ (or MAR) holds depends on what variables are included among those in **Z**. This property may be troublesome. One can show that MAR+ can hold either: a) after but not before a variable is added to **Z**; or b) before but not after a variable is added to **Z**. The same is true for MAR.

For age, for example, the sharply lower *P*-value in the first run may simply be due to age being positively associated with missingness if 30-day mortality is not controlled for, but not if it *is* controlled for. In contrast, the reason for a sometimes-missing variable X having a low *P*-value in the second run but not in the first may be harder to grasp. One explanation is that a negative (positive) association between missingness and X when 30-day mortality is controlled for can disappear when it is not controlled for, if mortality is associated positively (negatively) with X and positively with missingness.

Example 4. This example is based on data from 9580 patients collected retrospectively for the Study of Patients Intolerant of Converting Enzyme Inhibitors (SPICE) Registry, whose purpose was to provide information to examine the use of angiotensin-converting enzyme (ACE) inhibitors among patients with left ventricular systolic dysfunction.[24] The registry included patients from 105 study centers in eight countries. Of the 10 variables in our data set, $s = 8$ (listed in Table 4) are sometimes missing and $r = 2$ are never missing.

Both never-missing variables are dichotomous, so the number of $\mathbf{Z}$ cells is 2×2, or 4. One of these variables is geographical area, either North America or Europe. The other, which was the response variable, indicates the use or nonuse of ACE inhibitors.

As before, $Q_{\mathrm{MH}(3)}$ for X versus D was calculated using each sometimes-missing variable as X. Table 4 shows the eight resulting *P*-values and the signs of the association between X and D.

All but two of the *P*-values are significant at the 0.05 level. Those for age and New York Heart Association class are significant below the 0.00001 level. Greater missingness is associated with older patients, female sex and higher creatinine levels but with lower (better) New York Heart Association class and higher ejection fraction. One easily rejects MAR+ upon multiplying the lowest *P*-value by 8 through application of the Bonferroni inequality.

We note that, although the $Q_{\mathrm{MH}(3)}$ *P*-values in each of Tables 2–4 were obtainable with SAS[25] through PROC FREQ, the same was not true of the signs of the associations. To obtain the signs, we had to code detailed calculations.

Table 4 Results for the $s = 8$ sometimes-missing variables of Example 4

Variable	Type of variable	Percentage of data missing (%)	$Q_{\mathrm{MH}(3)}$ *P*-value	Sign of association
Age	Continuous	3	<0.00001	+
Ejection fraction	Continuous	14	0.001	+
Potassium	Continuous	9	0.03	−
Sodium	Continuous	9	0.10	−
New York Heart Association class (1, 2, 3 or 4, best to worst)	Ordered categorical	10	<0.00001	−
Creatinine (0, low; 1, high)	Dichotomous	10	0.02	+
Sex (0, female; 1, male)	Dichotomous	1	0.00002	−
Ischernic etiology (0, no; 1, yes)	Dichotomous	2	0.18	−

5 Statistical tests of MAR+ for $s \geq 2$ and $r_2 > 0$

We turn now to the most complicated case, where at least one never-missing variable in Z is continuous rather than categorical. As in Section 4, the aim is to test MAR+ by testing for independence between D and each X, conditional on $\mathbf{z}$ and conditional on X not being missing. That is, separately for each of the s sometimes-missing variables X, we wish to test whether X and D are independent (for fixed $\mathbf{z}$) among those patients for whom X is not missing.

Because of the greater complexities when $r_2 > 0$, it appears that there are even more ways to test independence than in Section 4 but also that any test that is chosen will require more assumptions than when $r_2 = 0$. As before, we put forth one test that appears reasonable and simple but briefly mention other possibilities as well.

With even one continuous variable in $\mathbf{Z}$, it is no longer suitable to have a separate cell for each $\mathbf{z}$ combination, because then each cell would (assuming perfect continuousness) contain only one patient. Thus, $Q_{\mathrm{MH}(3)}$ cannot be used. We therefore consider a proportional-odds model, [26] of the form

$$\log \frac{\Pr\{D \leq i|\mathbf{z}, x\}}{1 - \Pr\{D \leq i|\mathbf{z}, x\}} = b_i + b_{h0} + \sum_{j=r_1+1}^{r} b_{hj} z_j + bx$$

where i indexes the number of missing variables ($i = 0, 1, 2, \ldots, s-2$, except that some of these i-values will be absent if some of the s possible values of D fail to occur at all among the patients for whom X is not missing); h indexes the cells constructed from the r_1 categorical never-missing variables in $\mathbf{Z}$($h = 1, 2, \ldots, N$, where N denotes the total number of such cells, with $N = 1$ if $r_1 = 0$); and j indexes the r_2 continuous never-missing variables in $\mathbf{Z}$, with the z_j's denoting their values.

For the smallest i, the term b_i is to be omitted (thus, $b_0 = 0$ if $D = 0$ occurs). Note that h for a given patient is determined by $\mathbf{z}$ for that patient (more specifically, by the values of the patient's r_1 categorical never-missing variables in $\mathbf{Z}$). The parameters to be estimated for the model consist of $(s-2)$ or fewer b_i's, N b_{h0}'s, Nr_2 b_{hj}'s and b. The test of independence is done simply by testing the null hypothesis that b, the regression coefficient for X, is equal to 0.

Because of its parallelism condition among other things, the proportional-odds model is more restrictive and entails more assumptions than is true of the theoretical foundations of Section 4, in which all the variables in $\mathbf{Z}$ are categorical (ie, $r_2 = 0$). The greater restrictiveness stems from the model equation just described, which requires a certain linear form for the logit of $\Pr\{D \leq i|\mathbf{z}, x\}$ that is absent from the nonparametric approach of Section 4. Although the proportional-odds test could still be used even if $r_2 = 0$ [the right side of its model equation is then just $(b_{h0} + bx)$ if $s = 2$, eg], the added assumptions that it entails would usually seem to render it inferior to $Q_{\mathrm{MH}(3)}$ for $r_2 = 0$.

We alter the suppositions i) and ii) of Section 3. Of course, i), which refers to avoiding an excessive value of N, pertains no longer to all r never-missing variables of $\mathbf{Z}$ but only to the r_1 categorical never-missing variables. As for ii), it can be weakened in the sense that one need not rule out a sometimes-missing variable X that is unordered categorical.

The model equation described earlier is not general enough to cover such a variable, but can be modified to do so. One replaces bx with b_{*_e} for a patient whose X falls in the eth category (where b_{*_e} for the reference category is taken to be 0). One then tests the null hypothesis that all the b_{*_e}'s are 0 instead of testing $b = 0$.

Although for our Example 5 we will apply the proportional-odds model that we have put forth, we now mention, with little elaboration, some other approaches to independence testing. First, one could modify the right side of the model equation. One could *reduce* the number of parameters by, for example, replacing each b_{hj} with b_{0j}, that is, by dropping the dependence on h. Or one could use *more* terms and *more* parameters on the right side, such as by including terms in $z_j z_{j'} (j' \neq j)$ or z_j^2 in addition to those in z_j or by changing b to b_{h*}. (If b were replaced with b_{h*}, then the null hypothesis would be $b_{h*} = 0$ for all h rather than just $b = 0$.) One would probably be more inclined to use simpler parameterization on the right side of the model equation if N is high than if N is low, because of greater vulnerability to nonestimability problems with higher N.

Another approach to the independence testing would be to put X on the left side of the model equation and D on the right, rather than vice versa. One could argue, though, that it is more logical to predict D from X than X from D. A further drawback would be that, whereas predicting D from X requires just one type of model (proportional odds), prediction of X from D would require several types, with continuous, dichotomous, ordered categorical and unordered categorical X each needing a different type.

Finally, one might consider nonparametric tests of association between D and X along lines[27–29] that involve caliper matching on $\mathbf{Z}$, test statistics similar to Kendall's tau and theory of U-statistics. This approach has complexities but avoids some assumptions of the proportional-odds model.

Example 5. Our final example uses data from the multicenter Integrilin to Minimize Platelet Aggregation and Coronary Thrombosis-II (IMPACT-II) randomized, double-blind clinical trial of 4010 patients undergoing coronary intervention.[30] Our data set for these patients contains variables that were used for analyses dealing with prediction of contrast-induced nephropathy.[31] There are $s = 7$ sometimes-missing variables, shown in Tables 5, and $r = 4$ never-missing variables.

Table 5 Results for the $s = 7$ sometimes-missing variables of Example 5

Variable	Type of variable	Percentage of data missing (%)	P-value for $\hat{b}$	Sign of association
Logarithm of peak post-procedural creatinine level	Continuous	7	0.04	+
Logarithm of baseline creatinine level	Continuous	1	0.00167	+
Body surface area	Continuous	1	0.33	+
Ejection fraction	Continuous	19	0.91	−
Proteinuria[a]	Dichotomous	30	0.94	−
Diabetes[a]	Dichotomous	0.1	0.998	+
Hypertension[a]	Dichotomous	0.1	0.07	+

[a]Coded 1, condition present; 0, condition absent.

Of the latter, $r_1 = 3$ are categorical and $r_2 = 1$ is continuous. The continuous variable is age. The categorical variables are sex, the presence or absence of preprocedure treatment with a calcium-channel antagonist and the presence or absence of preprocedure treatment with an ACE inhibitor. All three of these are dichotomous, so the number of cells for the categorical never-missing variables in $\mathbf{Z}$ is $N = 2 \times 2 \times 2 = 8$.

We ran the proportional-odds model using each sometimes-missing variable as X. Each run estimated $N = 8$ b_{h0}'s, $Nr_2 = 8$ regression coefficients b_{hj} (all for age) and the b_i's (either b_1, b_2 or b_1, b_2, b_3, depending on the run). But the most important output of any run was, of course, $\hat{b}$ (the estimate of b, the regression coefficient for X) and its associated P-value.

For each of the seven sometimes-missing variables, Table 5 shows the P-value for $\hat{b}$ and the sign of $\hat{b}$. Hypertension is significant at the 0.07 level, peak creatinine at the 0.04 level and baseline creatinine at the 0.00167 level. For all three, the association with missingness is positive. Upon applying the Bonferroni inequality by multiplying 0.00167 by $s = 7$, one obtains an overall P-value slightly above 0.01, thus leading to rejection of MAR+ if one accepts the assumptions of the proportional-odds model.

6 Concluding remarks

The examples of this paper illustrate a range of situations where our techniques are applicable. Our methods provide comparatively simple and understandable ways for cautious investigators to better assess, in some cases, whether it is valid to assume MAR.

We note again that our techniques are not intended for a longitudinal study with missing data in a monotone response pattern, nor for any other situation where failure to observe one sometimes-missing variable logically precludes another sometimes-missing variable from being observed. In none of the examples of the paper are there cases where the missingness of one variable logically precludes the presence of another variable. In fact, upon checking all $s(s-1)$ ordered pairs of sometimes-missing variables in each example, we found only one pair where all patients who were missing the first variable were also missing the second. (Namely, in Example 3, all patients with missing data on prior bypass surgery were also unobserved for family history of coronary heart disease.)

We summarize and review some special cases. If there is only $s = 1$ sometimes-missing variable, then MAR and MAR+ are indeed equivalent (irrespective of any conditions), but our methods do not apply and no test statistic can be calculated. Otherwise, our tests of MAR+ can always be run, but are useful for assessing MAR only if Conditions 1 and 2 both hold. If there are $r = 0$ never-missing variables, then MAR+ is equivalent to MCAR, and to MAR as well if Conditions 1 and 2 both hold, in which case MCAR and MAR are the same so that our tests test for both.

We have presented five examples, all with real-world data. In all but Example 2, MAR+ was rejected, with a P-value of about 0.01 (Examples 1 and 5) or much lower (Examples 3 and 4). Because MAR+ implies MAR, failure to reject MAR+ provides some reason to believe that MAR is a valid assumption, although a small sample size, as in Example 2 ($n = 69$), limits any such assurance somewhat. Of course, larger sample sizes increase the chance of rejecting MAR+ if MAR+ does not hold, and of detecting

smaller deviations from MAR+. In Example 1, however, MAR+ was rejected despite a small sample size. The sample sizes in the last three examples were in the thousands, thus providing greater chance that minor violations of MAR+ would trigger rejection; however, the *P*-values for rejecting MAR+ were extremely low for Examples 3 and 4 (though not for Example 5). Remember that there is only one overall *P*-value for each example, obtained from a set of separate *P*-values through the Bonferroni inequality so as to prevent distortion from multiple significance tests.

What does one conclude if MAR+ fails? First, under certain conditions that are often reasonable to assume, failure of MAR+ implies failure of MAR (Proposition 2). The judgement of how closely these conditions apply will affect the extent to which one is confident or dubious about MAR. To assess the conditions, one should ideally try to find out what factors are causing or affecting the missingness. This may be difficult. But one can try to examine what types of patients tend to have missing data and why. Association between missingness and unhealthiness indicators whose direction is consistent (inconsistent) across indicators may suggest good (poor) conformity with the conditions.

Even if MAR fails, one does not generally know whether the effects of using a statistical technique that assumes MAR will be serious or minor. One study[15] found that effects can be substantial. Another one,[32] using only $s = 1$ sometimes-missing variable, found that consequences of MAR violations could range from negligible to serious, depending on circumstances. In any case, though, rejection of MAR+ should at least inject a note of caution regarding any approach that assumes MAR. One should not then discard the analyses from such an approach, but neither should one present them without suitable warnings if Conditions 1 and 2 seem plausible. A new approach that does not assume MAR, such as nonignorable modelling, might be tried in some cases.

In view of Proposition 2, one should be particularly skeptical about MAR if MAR+ is rejected and more missingness is associated consistently either with sicker or with healthier patients. Sicker patients unambiguously have more missing data than healthier ones in Example 1 (although it has only $s = 2$ sometimes-missing variables) and in Example 5 (where missingness is greater for sicker patients for each of the three variables whose *P*-values in Table 5 are <0.10).

Example 3 largely also follows the pattern of sicker patients having more missing data. There are seven sometimes-missing variables in Table 3 with either one or two *P*-values < 0.01. Greater missingness is associated with lower weight, lower systolic and diastolic blood pressure, older age, higher Killip class, diabetes and higher 30-day mortality. Because an earlier statistical analysis[33] of all 41021 patients in the GUSTO-I trial showed that lower systolic blood pressure and lower weight are independent predictors of 30-day mortality, it is not unreasonable to think that lower weight and lower blood pressure are indicative of poorer health in our data set for Example 3.

In addition, a separate analysis fit a logistic-regression model to predict death at 30 days using all patients, with missing data imputed with a single-imputation[34, pp. 69–70] method. The analysis found that the most significant predictor ($P < 0.0001$) was a variable that represented the number of variables that were missing and had to be imputed (our variable *D*). Thus, the patients with missing data were probably the sickest patients, or at least the ones most likely to die. Among variables other than *D* in this analysis, some were significant and some were not.

For Example 4, the remaining example where MAR+ was rejected, the results are more difficult to interpret, despite an overall P-value below the 0.0001 level. Some variables in Table 4 do show greater missingness associated with characteristics reflecting sicker patients. Two variables, however, stand out in exhibiting an association between healthier patients and more missingness: patients with higher ejection fraction and those with lower New York Heart Association class have more missing data. Thus, the validity of Conditions 1 and 2 (Proposition 2) is open to serious question, because missingness is not associated consistently either with sicker patients or with healthier ones. All of this means that the failure of MAR+ may not imply failure of MAR. The retrospective collection of the data for Example 4 might somehow cause peculiar results with regard to missing data, but it is not clear how that could happen.

In a logistic regression in which missing values were imputed through single imputation, Example 4, like Example 3, showed a highly significant association ($P < 0.0001$) between the response variable (use of ACE inhibitors in this case) and number of missing values. (The association was negative.) Except for sodium, ischemic etiology and New York Heart Association class, all other variables were very significant predictors of the response as well. Thus, the most significant sometimes-missing variables in the logistic-regression model are the same as those other than New York Heart Association class that show small $Q_{\mathrm{MH}(3)}$ P-values in Table 4 in the tests for MAR+.

If there is an indication, stemming from rejection of MAR+, that MAR fails (or may fail) to be satisfied, one has to decide how to do the statistical analysis. One could still use ordinary multiple imputation (which is generally better than single-imputation methods), but provide a cautionary warning that MAR may not hold and that results may (or may not) be seriously affected if it does not hold. More than one type of analysis may be helpful to see if the different results are roughly consistent. Although imputation through nonignorable modeling of some form is a possibility, its complexities may be a drawback. There are other possibilities.

In particular, if a situation involves prediction of one variable conditional on others, as through regression, should one consider running a complete-case analysis (ie, one that excludes all patients except those who have no missing values)? Because of their simplicity, complete-case analyses are often done, whether justifiably or not. Of course, they suffer from loss of efficiency, more so when the percentage of patients lacking complete data is larger. They are also in disfavor because of questionable validity (often a more serious problem than loss of efficiency).

Nevertheless, it is mathematically possible for complete-case analysis to be valid when MAR fails.[7, pp. 1229,1234] For instance, suppose that there are $s = 2$ sometimes-missing variables, X and Y, and $r = 1$ never-missing variables, Z, all three of which are dichotomous with possible values of 0 and 1, and suppose that z (representing eg, mortality within a given time period) is to be predicted from (x, y). For both $z = 0$ and $z = 1$, let $P(00|x, y) = 0.91 - 0.12x - 0.12y$, $P(01|x, y) = 0.04 + 0.02x + 0.06y$, $P(10|x, y) = 0.04 + 0.06x + 0.02y$ and $P(11|x, y) = 0.01 + 0.04x + 0.04y$. Then MAR fails (as do MAR+ and MCAR). But complete-case analysis is valid for predicting z, because missingness of X and Y does not depend on z and so the conditional distribution of z given x and y is the same for the complete cases as it is for all patients. Whether a situation like this would occur in practice, however, is naturally open to question.

In addition, when the patients with the most missing data are the sickest and the most likely to reach an endpoint such as stroke or death, then, in a complete-case analysis, one tends to remove patients with the most extreme values for a given variable. One thus reduces the variability of the variable and perhaps weakens its chance to be a significant predictor. Under these circumstances, it would be helpful to include a clear description of the types of patients with and without missing data to help clarify the populations whose data the analyses now deal with.

We did compare how different imputation methods and complete-case analysis worked for the data in Examples 3–5. The imputation methods included median substitution, single imputation and multiple imputation. As expected, highly significant predictors were relatively insensitive to missing data and to the imputation method. Also as expected, variables were more significant for all imputation methods, including median substitution, than for complete-case analysis. If MAR+ is violated, it is harder to decide which results are the most accurate. Variability of predictors may be underestimated if missingness is related too strictly to the sicker (or healthier) patients, and estimates of regression coefficients may be biased. An analysis that describes the patients with missing data versus those with little or no missing data should accompany all results to help provide some understanding of the type of population that the data represent. For example, data from a complete-case analysis may represent a very different population compared with the data that contain imputed values, depending on the percentage of missing values and whether MAR+ is violated.

There are limits to the ability of sophisticated statistical methods to correct for missing data. Unfortunately, it is not easy to determine before a study starts whether MAR or MAR+ will hold or the implications if either does not hold. Thus, one should consider a complementary mode of attack, namely, to take steps to try to prevent missing data in the first place. Little[8, p. 2622] advises that 'the most important step in dealing with missing data is to try to avoid it during the data-collection stage.' As Vach and Blettner[35, p. 2652] state in starting their discussion of strategies to cope with missingness (in the context of epidemiologic studies), 'The best advice is to minimize the possibility for missing values.' Even if it is not possible to avoid missing data completely, any statistical technique that is used to handle missingness will be less vulnerable to adverse statistical consequences the fewer missing values that exist. Often it may not be as difficult to reduce missingness in medical studies as it is in, for example, sample surveys, because in the former, investigators may have more control over conditions than in the latter.

Thus, approaches to act pre-emptively to diminish the amount of missing data in medical studies may deserve to receive more attention than they have heretofore attracted. Such approaches might include undertaking timely steps to retrieve missing data while and if such recovery is still possible; allowing somewhat imprecise data where the only alternative is no data at all; avoiding collection of data on variables of marginal value so that efforts can be concentrated on the more important variables; organizational initiatives involving such matters as accountability and incentives; and improvements regarding forms, along with increased computerization of data collection and data handling. In fact, some of the concepts of quality-control statisticians and of total quality management[36] might be usefully adapted to medical studies.

Acknowledgements

We wish to thank Anastasios A. Tsiatis and Gary G. Koch for their early helpful advice concerning our research. We also thank the referee for many useful comments.

References

1 Rubin DB. *Multiple imputation for nonresponse in surveys*. Wiley, 1987.
2 Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**: 473–89.
3 Barnard J, Rubin DB, Schenker N. Multiple imputation methods. In Armitage P, Colton T, eds. *Encyclopedia of biostatistics*. Wiley, 1998: 2772–80.
4 Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–92.
5 Little RJA, Rubin DB. *Statistical analysis with missing data*. Wiley, 1987.
6 Little RJA, Rubin DB. *Statistical analysis with missing data*. 2nd edn. Wiley, 2002.
7 Little RJA. Regression with missing X's: a review. *Journal of the American Statistical Association* 1992; **87**: 1227–37.
8 Little RJ. Missing data. In Armitage P, Colton T, eds. *Encyclopedia of biostatistics*. Wiley, 1998: 2622–35.
9 Schafer JL. *Analysis of incomplete multivariate data*. Chapman and Hall, 1997.
10 Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**: 1112–21.
11 Little RJA. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* 1988; **83**: 1198–202.
12 Park T, Davis CS. A test of the missing data mechanism for repeated categorical data. *Biometrics* 1993; **49**: 631–8.
13 Chen HY, Little R. A test of missing completely at random for generalized estimating equations with missing data. *Biometrika* 1999; **86**: 1–13.
14 Qu A, Song PX-K. Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika* 2002; **89**: 841–50.
15 Vach W, Blettner M. Logistic regression with incompletely observed categorical covariates – investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine* 1995; **14**: 1315–29.
16 van Elteren P. On the combination of independent two sample tests of Wilcoxon. *Bulletin of the International Statistical Institute* 1960; **37**: 351–61.
17 Lehmann EL. *Nonparametrics: statistical methods based on ranks*. Holden-Day, 1975.
18 Fuchs C. Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association* 1982; **77**: 270–8.
19 Landis JR, Sharp TJ, Kuritz SJ, Koch GG. Mantel-Haenszel methods. In Armitage P, Colton T, eds. *Encyclopedia of biostatistics*. Wiley, 1998: 2378–91.
20 Koch GG, Gillings DB, Stokes ME. Biostatistical implications of design, sampling, and measurement to health science data analysis. *Annual Review of Public Health* 1980; **1**: 163–225.
21 Stokes ME, Davis CS, Koch GG. *Categorical data analysis using the SAS system*. SAS Institute, 1995.
22 Little RJA, Schluchter MD. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* 1985; **72**: 497–512.
23 Holmes DR Jr, Bates ER, Kleiman NS *et al*. Contemporary reperfusion therapy for cardiogenic shock: the GUSTO-I trial experience. *Journal of the American College of Cardiology* 1995; **26**: 668–74.
24 Bart BA, Ertl G, Held P *et al*. Contemporary management of patients with left ventricular systolic dysfunction: results from the study of patients intolerant of converting enzyme inhibitors (SPICE) registry. *European Heart Journal* 1999; **20**: 1182–90.
25 SAS Institute Inc. *SAS/STAT user's guide*, Version 8. SAS Institute, 1999.
26 Agresti A. *Categorical data analysis*. Wiley, 1990.

27 Carr GJ, Hafner KB, Koch GG. Analysis of rank measures of association for ordinal data from longitudinal studies. *Journal of the American Statistical Association* 1989; **84**: 797–804.
28 Quade D. Nonparametric partial correlation. In Blalock HM Jr, ed. *Measurement in the social sciences: theories and strategies*. Aldine, 1974: 369–98.
29 Quade D. Nonparametric analysis of covariance by matching. *Biometrics* 1982; **38**: 597–611.
30 The IMPACT-II Investigators. Randomized placebo-controlled trial of effect of eptifibatide on complications of percutaneous coronary intervention: IMPACT-II. *Lancet* 1997; **349**: 1422–8.
31 Thel MC, Davidson CJ, Aguirre FV *et al*. What predicts contrast nephropathy after coronary intervention? Insights from the IMPACT II trial. *Journal of the American College of Cardiology* 1997; **29**: 500A.
32 Collins LM, Schafer JL, Kam C-M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**: 330–51.
33 Lee KL, Woodlief LH, Topol EJ *et al*. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction: results from an international trial of 41,021 patients. *Circulation* 1995; **91**: 1659–68.
34 Harrell FE Jr. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, 2001.
35 Vach W, Blettner M. Missing data in epidemiological studies. In Armitage P, Colton T, eds. *Encyclopedia of biostatistics*. Wiley, 1998: 2641–54.
36 Roberts HV. Total quality management. In Kotz S, Read CB, Banks DL, eds. *Encyclopedia of statistical sciences*. Update Volume 1. Wiley, 1997: 525–8.

Appendix: proof of Proposition 2 for $s = 3$

For $s = 3$, let the three sometimes-missing variables be W, X and Y. Define the variables T, U and V to be 1 if W, X and Y, respectively, are missing, and 0 if W, X and Y are not missing. Let $\mathbf{m} = (t, u, v)$. Let $P(tuv|w, x, y)$ denote the probability, conditional on w, x and y (as well as $\mathbf{z}$), that $\mathbf{M} = (t, u, v)$.

Under MAR, $P(tuv|w, x, y)$ must satisfy

$$P(111|w,x,y) = f_{111}$$
$$P(011|w,x,y) = f_{011}(w), \quad P(101|w,x,y) = f_{101}(x), \quad P(110|w,x,y) = f_{110}(y)$$
$$P(001|w,x,y) = f_{001}(w,x), \quad P(010|w,x,y) = f_{010}(w,y), \quad P(100|w,x,y) = f_{100}(x,y)$$
$$P(000|w,x,y) = f_{000}(w,x,y)$$

These eight probabilities have to add up to 1 for all (w, x, y).

MAR+ is, of course, more stringent. Each of the eight probabilities must be a constant (for any fixed $\mathbf{z}$).

The proof proceeds as follows. With $s = 3$, there are $(2^s - 2) = 6$ probabilities that are the focus of Condition 1. One of them is

$$P(11 \cdot |w,x,y) = f_{111} + f_{110}(y)$$

where the notation is like that used earlier. From this expression and from Condition 1, it follows that $f_{110}(y)$ is nondecreasing in y. One can show similarly that $f_{011}(w)$ is nondecreasing in w and $f_{101}(x)$ is nondecreasing in x.

For $s = 3$, there are $\binom{s}{2}2^{s-2} = 6$ conditional probabilities that are the focus of Condition 2. They can be denoted by $g_{1**}, g_{0**}, g_{*1*}, g_{*0*}, g_{**1}$ and g_{**0}, where each one can be written as a function of (w, x, y). The two asterisks in the subscripts identify the two variables referred to as X' and X'', with the first asterisk corresponding to X' and the second to X''. The 1 or 0 subscript indicates conditioning on the remaining variable being missing (1) or present (0). For example

$$g_{1**}(w, x, y) = \frac{P(110|w, x, y)}{P(110|w, x, y) + P(101|w, x, y)} = \frac{f_{110}(y)}{f_{110}(y) + f_{101}(x)}$$

By Condition 2, $g_{1**}(w, x, y)$ is nondecreasing in x and nonincreasing in y, from which it follows that $f_{101}(x)$ is nonincreasing in x and $f_{110}(y)$ is nonincreasing in y. In a similar way, one can use $g_{*1*}(w, x, y)$ or $g_{**1}(w, x, y)$ to show that $f_{011}(w)$ is nonincreasing in w.

From the results so far, it follows that $f_{011}(w), f_{101}(x)$ and $f_{110}(y)$ are all constants. Now note that, for example

$$P(\cdot 1 \cdot |w, x, y) = f_{111} + f_{011}(w) + f_{110}(y) + f_{010}(w, y)$$

Because $f_{011}(w)$ and $f_{110}(y)$ are both constant, Condition 1 implies that $f_{010}(w, y)$ is nondecreasing both in w (for any fixed y) and in y (for any fixed w).

By Condition 2,

$$g_{**0}(w, x, y) = \frac{P(100|w, x, y)}{P(100|w, x, y) + P(010|w, x, y)} = \frac{f_{100}(x, y)}{f_{100}(x, y) + f_{010}(w, y)}$$

is nondecreasing in w for any fixed x and y, so $f_{010}(w, y)$ is nonincreasing in w (for any fixed y). Also by Condition 2,

$$g_{0**}(w, x, y) = \frac{P(010|w, x, y)}{P(010|w, x, y) + P(001|w, x, y)} = \frac{f_{010}(w, y)}{f_{010}(w, y) + f_{001}(w, x)}$$

is nonincreasing in y for any fixed w and x, so $f_{010}(w, y)$ is nonincreasing in y (for any fixed w).

Thus, $f_{010}(w, y)$ is constant. A similar argument establishes that $f_{100}(x, y)$ and $f_{001}(w, x)$ are each constant.

Finally, $f_{000}(w, x, y)$ must be constant, as the other seven probabilities are all constant. Thus, MAR+ is satisfied.

As indicated earlier, the proof for general s works in basically the same way as the proof for $s = 3$.