Statistical Methods in Medical Research 1999; **8**: 3–15

Multiple imputation: a primer

Joseph L Schafer Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania, USA

In recent years, multiple imputation has emerged as a convenient and flexible paradigm for analysing data with missing values. Essential features of multiple imputation are reviewed, with answers to frequently asked questions about using the method in practice.

1 Introduction

Imputation, the practice of 'filling in' missing data with plausible values, has long been recognized as an attractive approach to analysing incomplete data. For decades, survey statisticians have been imputing large databases by often elaborate means.[1] From an operational standpoint, imputation solves the missing-data problem at the outset, enabling the analyst to proceed without further hindrance. From a statistical standpoint, however, a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests, as documented by Little and Rubin[2] and others.

The question of how to obtain valid inferences from imputed data was addressed by Rubin[3] in his book on multiple imputation (MI). MI is a Monte Carlo technique in which the missing values are replaced by $m > 1$ simulated versions, where m is typically small (say, 3–10). In Rubin's method for 'repeated imputation' inference, each of the simulated complete datasets is analysed by standard methods, and the results are later combined to produce estimates and confidence intervals that incorporate missing-data uncertainty. Rubin's text addresses potential uses of MI primarily for large public-use data files from sample surveys and censuses. With the advent of new computational methods and software for creating MI's, however, the technique has become increasingly attractive for researchers in the biomedical, behavioural, and social sciences whose investigations are hindered by missing data. These methods are documented in a recent book by Schafer[4] on incomplete multivariate data.

MI is not the only principled method for handling missing values, nor is it necessarily the best for any given problem. In some cases, good estimates can be obtained through a weighted estimation procedure (see, e.g. Little[5] and Robins *et al.*[6]). In fully parametric models, maximum-likelihood estimates can often be calculated directly from the incomplete data by specialized numerical methods, such as the EM algorithm.[7] The estimates obtained through such procedures may be somewhat more efficient than those from MI, because they involve no simulation. Given sufficient time and resources, one could perhaps derive a better statistical procedure than MI for any

Address for correspondence: JL Schafer, Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802-6202, USA. E-mail: jls@stat.psu.edu

 0962-2802(99)SM176RA

particular problem. In real-life applications, however, where missing data are a nuisance rather than a major focus of scientific enquiry, a readily available, approximate solution with good properties can be preferable to one that is more efficient but problem-specific and complicated to implement. MI is not the only tool available, but it is a handy one and a valuable addition to any data analyst's toolkit.

In the remainder of this article I provide an overview of the MI paradigm and discuss practical issues in the form of answers to frequently asked questions. Techniques and software for creating MIs are reviewed, followed by an example of MI in a simple categorical data problem.

2 The MI paradigm

Rubin[3] presented the following method for repeated-imputation inference. Let Q denote a generic scalar quantity to be estimated, such as a mean, correlation, regression coefficient, or odds ratio. Let Y denote the intended data, part of which is observed (Y_{obs}) and part of which is missing (Y_{mis}). Let $\hat{Q} = \hat{Q}(Y_{obs}, Y_{mis})$ denote the statistic that would be used to estimate Q if complete data were available, and let $U = U(Y_{obs}, Y_{mis})$ be its squared standard error. We must assume that with complete data, tests and intervals based on the normal approximation

$$(\hat{Q} - Q)/\sqrt{U} \sim N(0,1) \tag{2.1}$$

would be appropriate. For this reason, it may be helpful to transform the estimand to a scale for which (2.1) works well, e.g. by taking the log of an odds ratio. A method appropriate for small-sample problems where (2.1) is replaced by a Student's t-distribution is discussed by Barnard and Rubin.[8]

In the absence of Y_{mis}, suppose that we have $m > 1$ independent simulated versions or imputations $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(m)}$. From these we calculate the imputed-data estimates $\hat{Q}^{(\ell)} = \hat{Q}(Y_{obs}, Y_{mis}^{(\ell)})$ along with their estimated variances $U^{(\ell)} = U(Y_{obs}, Y_{mis}^{(\ell)})$, $\ell = 1, \ldots, m$. The overall estimate of Q is simply the average

$$\bar{Q} = m^{-1} \sum \hat{Q}^{(\ell)} \tag{2.2}$$

To obtain a standard error for $\bar{Q}$, one must calculate the between-imputation variance $B = (m-1)^{-1} \sum (\hat{Q}^{(\ell)} - \bar{Q})^2$ and the within-imputation variance $\bar{U} = m^{-1} \sum U^{(\ell)}$. The estimated total variance is

$$T = (1 + m^{-1})B + \bar{U} \tag{2.3}$$

and tests and confidence intervals are based on a Student's t-approximation

$$(\bar{Q} - Q)/\sqrt{T} \sim t_\nu \tag{2.4}$$

with degrees of freedom

$$\nu = (m-1)\left[1 + \frac{\bar{U}}{(1+m^{-1})B}\right]^2$$

Notice that if Y_{mis} carried no information about Q, then the imputed-data estimates $\hat{Q}^{(\ell)}$ would be identical and T would reduce to $\bar{U}$. Therefore, $r = (1 + m^{-1})B/\bar{U}$ measures the relative increase in variance due to missing data, and the rate of missing information in the system is approximately $\lambda = r/(1 + r)$. A more refined estimate of this fraction, obtained by comparing the spread of (2.1) to (2.4), is

$$\lambda = \frac{r + 2/(\nu + 3)}{1 + r} \tag{2.5}$$

Inferential questions that cannot be cast in terms of a one-dimensional estimand (e.g. goodness-of-fit tests) can be handled through multivariate generalizations of this rule.[9–11]

The great virtues of MI are its simplicity and its generality. The user may analyse the data by virtually any technique that would be appropriate if the data were complete. The validity of the method, however, hinges on how the imputations $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(m)}$ were generated. Clearly it is not possible to obtain valid inferences in general if imputations are created arbitrarily. The imputations should, on average, give reasonable predictions for the missing data, and the variability among them must reflect an appropriate degree of uncertainty. Rubin[3] provides technical conditions under which repeated-imputation method leads to frequency-valid answers. An imputation method which satisfies these conditions is said to be 'proper'. These conditions, like many frequentist criteria, are useful for evaluating the properties of a given method but provide little guidance for one seeking to create such a method in practice. For this reason, Rubin recommends that imputations be created through Bayesian arguments: specify a parametric model for the complete data (and, if necessary, a model for the mechanism by which data become missing), apply a prior distribution to the unknown model parameters, and simulate m independent draws from the conditional distribution of Y_{mis} given Y_{obs} by Bayes' theorem. In simple problems, the computations necessary for creating MI's can be performed explicitly through formulas. In nontrivial applications, however, special computational techniques such as Markov chain Monte Carlo, to be described in Section 4, must be applied.

When imputations are created under Bayesian arguments, Rubin's repeated-imputation method has a natural interpretation as an approximate Bayesian inference for Q based on the observed data. Suppose $\hat{Q}$ and U can be regarded as an approximate complete-data posterior mean and variance for Q, $\hat{Q} = E(Q \mid Y_{obs}, Y_{mis})$ and $U = V(Q \mid Y_{obs}, Y_{mis})$. Then (2.2) approximates the actual posterior mean

$$E(Q \mid Y_{obs}) = E(\hat{Q} \mid Y_{obs})$$

and (2.3) approximates the posterior variance

$$V(Q \mid Y_{obs}) = V(\hat{Q} \mid Y_{obs}) + E(U \mid Y_{obs})$$

The term $m^{-1}B$ in (2.3) and the use of t_ν rather than a normal distribution widen the resulting interval estimates to account for simulation error incurred by using $m < \infty$. Unless the fraction of missing information λ is unduly large, the widening effect is not substantial, and MI inferences are quite efficient even when m is small.

A key feature of MI is that the imputation phase is operationally distinct from subsequent analyses. The imputations $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(m)}$ may be created by one person or organization and the analyses carried out by another. This raises the possibility that the statistical model or assumptions used to create the imputed datasets may be somehow incompatible with those used to analyse them. The behaviour of repeated-imputation inference when the imputer's and analyst's models differ has been investigated by Fay,[12] Meng[13] and Rubin.[14] When the imputer's model is more general (i.e. makes fewer assumptions) than the analyst's, then MI leads to valid inferences with perhaps some loss of power, because the additional generality may add extra variation among the imputes $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(m)}$. When the imputer makes more assumptions than the analyst – and the extra assumptions are plausible – then the MI estimate $\bar{Q}$ may become more precise than any estimate derived from the observed data and analyst's model alone, a property that Rubin[14] calls 'superefficiency'. In such cases, MI intervals tend to be narrower than intervals derived purely from the analyst's model, and they also tend to be conservative with higher-than-nominal coverage probability, as shown by the theoretical results of Meng.[13]

The only serious danger of inconsistency arises when the imputer makes more assumptions than the analyst and these additional assumptions are unwarranted. For example, consider a situation where a variable is imputed under a no-interactions regression model and the analyst subsequently looks for evidence of interactions; if interactions are present, then the MI estimates of them will be biased toward null values. In practice, this means that an imputation model should reasonably preserve those distributional features (e.g. associations) that will be the subject of future analyses. Above all, the processes of imputation and analysis should be guided by common-sense. For example, suppose that variables with skewed, truncated, or heavy-tailed distributions are, for the sake of convenience, imputed under an assumption of joint normality. Analyses that depend primarily on means, variances, and covariances, such as regression or principal-component methods, should perform reasonably well even though the imputer's model is rather simplistic. That is, the coverage of the repeated-imputation intervals will tend to be no worse (and may actually be better) than those of the same procedure performed on the data without missing values; see, e.g. the simulation results of Graham and Schafer.[15] On the other hand, common-sense would suggest that the same imputations ought not be used for estimation of fifth or 95th percentiles, or other analyses sensitive to tail behaviour and other non-normal features.

3 Answers to frequently asked questions

This section addresses some common questions and concerns regarding the use of MI in practice. Questions of special relevance to producers of public-use databases are addressed by Rubin.[14]

Removing incomplete cases is so much easier than multiple imputation; why can't I just do that?

The shortcomings of various case-deletion strategies have been well documented.[2] If the discarded cases form a representative and relatively small portion of the entire dataset, then case deletion may indeed be a reasonable approach. However, case deletion leads to valid inferences in general only when missing data are missing completely at random (MCAR) in the sense that the probabilities of response do not depend on any data values observed or missing.[2] In other words, case deletion implicitly assumes that the discarded cases are like a random subsample. When the discarded cases differ systematically from the rest, estimates may be seriously biased. Moreover, in multivariate problems, case deletion often results in a large portion of the data being discarded and an unacceptable loss of power.

Why can't I just impute once?

Again, if the proportion of missing values is small, then single imputation may be quite reasonable. Without special corrective measures (e.g. the methods of Schafer and Schenker[16]), single-imputation inference tends to overstate precision because it omits the between-imputation component of variability. When the rate of missing information is small (say, less than 5%) then single-imputation inferences for a scalar estimand may be fairly accurate. For joint inferences about multiple parameters, however, even small rates of missing information may seriously impair a single-imputation procedure. In modern computing environments, the effort required to produce and analyse a multiply-imputed dataset is often not substantially greater than what is required for good single imputation.

How many imputations do I need?

Rubin[3] shows that the relative efficiency of an estimate based on m imputations to one based on an infinite number of them is approximately $(1 + \lambda/m)^{-1}$, where λ is the rate of missing information. With 50% missing information, an estimate based on $m = 5$ imputations has a standard deviation that is only about 5% wider than one based on $m = \infty$ because $\sqrt{1 + 0.5/5} = 1.049$. Unless rates of missing information are unusually high, there tends to be little or no practical benefit to using more than five to ten imputations.

Is multiple imputation like EM?

MI bears a close resemblance to the EM algorithm and other computational methods for calculating maximum-likelihood estimates based on the observed data alone. These methods summarize a likelihood function which has been averaged over a predictive distribution for the missing values. MI performs this same type of averaging by Monte Carlo rather than by numerical methods. In large samples, when the imputer's and analyst's models agree (i.e. are 'congenial' in the sense defined by Meng[13]), inferences obtained by MI with sufficiently large m will be nearly the same as those obtained by direct maximization of the likelihood. In smaller samples MI inferences may have better properties, because they are in effect approximating the observed-data posterior density by a finite mixture of normal densities rather than a single normal density, improving one's ability to capture non-normal features such as skewness or multiple modes.

Is multiple imputation related to MCMC and other simulation methods?

Markov chain Monte Carlo (MCMC) is a collection of methods for simulating random draws from nonstandard distributions via Markov chains. MCMC is one of the primary methods for generating MI's in nontrivial problems.[4] In much of the existing literature on MCMC – see the chapters of Gilks *et al*.[17] and their references, for example – MCMC is used for parameter simulation, for creating a large number of (typically dependent) random draws of parameters from Bayesian posterior distributions under complicated parametric models. In MI-related applications, however, MCMC is used to create a small number of independent draws of Y_{mis} from a predictive distribution, and these draws are then used for repeated-imputation inference. In many cases it is possible to conduct an analysis either by parameter simulation or by multiple imputation. Parameter simulation tends to work well when interest is confined to small number of well-defined parameters, whereas multiple imputation is more attractive for exploratory or multipurpose analyses involving a large number of estimands. Generating and storing $m = 10$ versions of Y_{mis} is often more efficient than generating and storing the hundreds or thousands of dependent draws that would be required to achieve a comparable degree of precision through parameter simulation.

What happens when the nonresponse is nonignorable?

Most of the techniques presently available for creating MI's assume that the nonresponse is 'ignorable' as defined by Rubin.[3] That is, they assume that missing data are missing at random (MAR) in the sense that the probability that an observation is missing may depend on observed values but not missing ones.[2] The MAR assumption is mathematically convenient because it allows one to eschew an explicit probability model for nonresponse. In some applications, however, MAR may seem artificial or implausible. With attrition in a longitudinal study, for example, it is possible that subjects drop out for reasons related to current data values. It is important to note that the MI paradigm does not require or assume that nonresponse is ignorable. Imputations may in principle be created under any kind of model for the missing-data mechanism, and the repeated-imputation method of Section 2 will produce valid answers under that mechanism. General techniques for creating MI's under alternative nonignorable models is an important area for future development.

Isn't multiple imputation just making up data?

When MI is presented to a new audience, some may view it as a kind of statistical alchemy in which information is somehow invented or created out of nothing. This objection is quite valid for single-imputation methods, which treat imputed values no differently from observed ones. MI, however, is nothing more than a device for representing missing-data uncertainty. Information is not being invented with MI any more than with EM or other well accepted likelihood-based methods, which average over a predictive distribution for Y_{mis} by numerical techniques rather than by simulation.

4 Techniques and software

4.1 MI from parametric Bayesian models

Rubin[3] describes methods for generating MIs by parametric Bayesian models in relatively simple problems. Despite their simplicity, these examples illustrate the basic principles of Bayesian imputation and lend insight into methods for more complicated and realistic problems.

Consider, for example, a univariate sample $Y = (y_1, \ldots, y_n)$ where the first a values $Y_{obs} = (y_1, \ldots, y_a)$ are seen and the remaining values $Y_{mis} = (y_{a+1}, \ldots, y_n)$ are missing at random. How does one create multiple imputations under the independent normal model $y_i \sim N(\mu, \psi)$, $i = 1, \ldots, n$ when $\theta = (\mu, \psi)$ is unknown? Under the standard noninformative prior $P(\theta) \propto \psi^{-1}$, it is straightforward to show that the observed-data posterior distribution of θ is

$$\mu \mid \psi, Y_{obs} \sim N(\bar{y}_{obs}, a^{-1}\psi)$$
$$\psi \mid Y_{obs} \sim (a-1)S_{obs}^2/\chi_{a-1}^2$$

where $\bar{y}_{obs} = a^{-1}\sum_{i=1}^{a} y_i$, $S_{obs}^2 = (a-1)^{-1}\sum_{i=1}^{a}(y_i - \bar{y}_{obs})^2$, and χ_{a-1}^2 denotes a chi-square variate with $a-1$ degrees of freedom. To create an imputation $Y_{obs}^{(\ell)} = (y_{a+1}^{(\ell)}, \ldots, y_n^{(\ell)})$, one would generate a random variance $\psi^{(\ell)} \sim (a-1)S_{obs}^2/\chi_{a-1}^2$ followed by a random mean $\mu^{(\ell)} \sim N(\bar{y}_{obs}, a^{-1}\psi^{(\ell)})$, and then draw $y_i^{(\ell)} \sim N(\mu^{(\ell)}, \psi^{(\ell)})$ independently for $i = a+1, \ldots, n$. Repeating the procedure for $\ell = 2, \ldots, m$ results in m proper imputations for Y_{mis}.

More generally, suppose that $Y = (Y_{obs}, Y_{mis})$ follows a parametric model $P(Y \mid \theta)$ where θ has a prior distribution and Y_{mis} is ignorably missing. Because

$$P(Y_{mis} \mid Y_{obs}) = \int P(Y_{mis} \mid Y_{obs}, \theta)\, P(\theta \mid Y_{obs})\, \mathrm{d}\theta$$

an imputation for Y_{mis} can be created by first simulating a random draw of the unknown parameters from their observed-data posterior

$$\theta^* \sim P(\theta \mid Y_{obs}) \tag{4.1}$$

followed by a random draw of the missing values from their conditional predictive distribution

$$Y_{mis}^* \sim P(Y_{mis} \mid Y_{obs}, \theta^*) \tag{4.2}$$

For many common models, (4.2) is straightforward but (4.1) is not. The observed-data posterior is typically not a standard distribution which can be easily simulated. Rubin[3] mentions a few general strategies for approximating draws from (4.1), including large-sample normal approximations and importance resampling. Soon after his book was published, however, simpler methods became available through the development of MCMC.

In MCMC, one creates a Markov chain with a desired stationary distribution. Overviews of popular MCMC methods, including Gibbs sampling and the Metropolis–Hastings algorithm, are provided by Gilks *et al*.[17] One MCMC method ideally suited to missing-data problems is the data augmentation algorithm of Tanner and Wong.[18]

Consider an iterative, two-step process in which we alternately sample missing values from their conditional predictive distribution $Y_{mis}^{(t)} \sim P(Y_{mis} \mid Y_{obs}, \theta^{(t-1)})$ and then sample unknown parameters from a simulated complete-data posterior $\theta^{(t)} \sim P(\theta \mid Y_{obs}, Y_{mis}^{(t)})$. Given an initial value $\theta^{(0)}$, this defines a Markov chain $\{(Y_{mis}^{(t)}, \theta^{(t)}), t = 1, 2, \ldots\}$ which, under quite general conditions, converges to the stationary distribution $P(Y_{mis}, \theta \mid Y_{obs})$. Executing these steps a large number of times eventually produces a draw of θ from its observed data posterior (4.1) and a draw of Y_{mis} from $P(Y_{mis} \mid Y_{obs})$, the distribution from which MIs are generated. In many cases, the second step of data augmentation $\theta^{(t)} \sim P(\theta \mid Y_{obs}, Y_{mis}^{(t)})$ is straightforward. In more complicated situations, this step is intractable and may be replaced by one or more cycles of another MCMC algorithm that converges to $P(\theta \mid Y_{obs}, Y_{mis}^{(t)})$.

MCMC provides a flexible set of tools for creating MIs from parametric models. MCMC methods for basic models for continuous, categorical, and mixed multivariate data are described by Schafer,[4] along with data examples and practical advice. Extensions to models with more complicated structure, such as clustering and repeated measurements, are also available.[19] These methods have been implemented by the author as functions in S-PLUS[20] and are available from `http://www.stat.psu.edu/~jls/misoftwa.html`. Some of these functions are incorporated into a missing-data module scheduled to be released with S-PLUS Version 5 in 1999.

A missing-data module in the current version of SPSS[21] performs maximum-likelihood estimation of means and covariances from an incomplete data matrix. The module also contains routines for predicted-mean and random imputation of missing values. By executing the random imputation procedure M times, it is possible to create multiple draws of the missing data. These multiple imputations are not proper, however, because the step (4.1) of sampling parameters from their observed-data posterior distribution is omitted.

4.2 Nonparametric methods

Consider again the univariate sample $y_1, \ldots, y_n$ where the first $a < n$ values are observed and the remaining $n - a$ values are missing. Is it possible to generate proper imputations for $Y_{mis} = (y_{a+1}, \ldots, y_n)$ with minimal distributional assumptions for $Y = (y_1, \ldots, y_n)$? Rubin[3] describes a simple method called the approximate Bayesian bootstrap (ABB) in which one creates: (a) a new pool of respondents Y_{obs}^* by sampling a values from $Y_{obs} = (y_1, \ldots, y_a)$ with replacement, and then (b) a set of imputed data Y_{mis}^* by sampling $n - a$ values from Y_{obs}^*, again with replacement. The method, which is most appropriate for large samples, produces approximate draws from $P(Y_{mis} \mid Y_{obs})$ under a multinomial model with categories corresponding to the distinct values seen in Y_{obs}. The resampling of Y_{obs}^* from Y_{obs} approximates a draw of the multinomial probabilities from their observed-data posterior (4.1) under a Dirichlet prior (see, for example, chapter 7 of Schafer[4]).

Now suppose that covariates $X = (X_1, \ldots, X_p)$ are available for each respondent and nonrespondent. The ABB can be extended in a variety of ways to incorporate the additional information provided by X. If the covariates are discrete and the respondent sample size is sufficiently large, it may be possible to partition the sample into cells corresponding to unique patterns of $(X_1, \ldots, X_p)$ and carry out the ABB procedure within each cell. With continuous covariates or large p, this strategy tends

to be ineffective because the observed data become too sparse. Using an idea from Rosenbaum and Rubin,[22] Lavori *et al.*[23] suggest defining response indicators $R = (r_1, \ldots, r_n)$, where $r_i = 1$ if unit i responded and $r_i = 0$ otherwise, and modelling the response propensities $\pi_i = P(r_i = 1)$ by logistic regression on the covariates X. The sample may then be partitioned into cells defined by coarse grouping (e.g. quintiles) of the fitted propensities $\hat{\pi}_i$ and the ABB procedure performed within each cell. It is possible to show that this strategy produces valid inferences about quantities pertaining to the distribution of Y when probabilities of missingness depend on X; grouping by response propensity effectively eliminates distortions that arise when respondents and nonrespondents differ in their X-distributions. This approach to MI has been implemented in a new commercial software product called Solas.[24]

It is important to note that the imputations produced by Solas are effective for analyses pertaining to the distribution of Y, but they are not appropriate in general for analyses involving relationships between Y and the covariates X. Consider a hypothetical covariate X_j that is highly correlated with the response Y but unrelated to the missingness indicators R. Imputed values for Y_{mis} will bear no relationship to X_j because that variable has no influence in the logistic regression model, and an MI-based estimate of the correlation between X_j and Y will be biased toward zero. The response-propensity ABB is unable to preserve important features of the joint distribution of X and Y. Partially parametric strategies for MI which can preserve these features are discussed by Schenker and Taylor.[25]

4.3 Software for repeated-imputation inference

The method of repeated-imputation inference for a scalar estimand described in Section 2 and its multivariate generalization[8] require only simple arithmetic and access to quantiles and tail probabilities of Student's t- and F-distributions. The computations are easily carried out in many statistical software packages. John Barnard has produced generic routines for repeated-imputation inference in Stata,[26] which are available at `http://www.stat.harvard.edu/~barnard/`. Functions incorporating MI inference are also forthcoming in S-PLUS Version 5.

5 Example

The data in Table 1, previously published by Marascuilo *et al.*[27] come from the Risk and Youth: Smoking Project, a school-based anti-smoking intervention programme. Pupils in the experimental-treatment group received an instructional programme on tobacco, whereas those in the control group received instructional materials on general science. Subjects were in the sixth and eighth grades. Pretest and post-test questionnaires included the statement, 'The new low-tar cigarettes aren't going to hurt me', with possible responses 'I agree' and 'I disagree'. Missing values arose at both occasions, due presumably to absenteeism, failure to complete the questionnaire, etc.

With complete data, changes in a binary response over time are commonly analysed using McNemar's test.[28] Consider a 2×2 table where n subjects are classified by yes/no answers at two occasions, Y_1 and Y_2. Let x_{ij} denote the observed count of subjects with response pattern $(Y_1 = i,\ Y_2 = j)$ and π_{ij} the probability of this pattern for

Table 1 Responses to 'The new low-tar cigarettes aren't going to hurt me' before (*B*) and after (*A*) intervention by treatment group and grade level

	Grade 6			Grade 8		
	Disagree_A	Agree_A	Missing_A	Disagree_A	Agree_A	Missing_A
Experimental group						
Disagree_B	61	18	4	91	12	9
Agree_B	55	70	14	35	19	3
Missing_B	12	20	—	28	9	—
Control group						
Disagree_B	69	16	7	100	23	11
Agree_B	13	18	1	26	34	0
Missing_B	24	12	—	37	31	—

Source: Marascuilo *et al*.[27]

$i = 1, 2$, $j = 1, 2$. The null hypothesis of no average change over time, $P(Y_1 = 1) = P(Y_2 = 1)$, which is equivalent to the hypothesis of symmetry, $\pi_{12} = \pi_{21}$, may be tested against the two-sided alternative by comparing McNemar' statistic $Z^2 = (x_{12} - x_{21})^2/(x_{12} + x_{21})$ to a χ^2_1 distribution. Changes over time are commonly expressed in terms of the difference $\Delta = \pi_{12} - \pi_{21}$, estimated by $\hat{\Delta} = \hat{\pi}_{12} - \hat{\pi}_{21}$ with standard error

$$\hat{V}^{1/2}(\hat{\Delta}) = \sqrt{n^{-1}[\hat{\pi}_{12}(1 - \hat{\pi}_{12}) + \hat{\pi}_{21}(1 - \hat{\pi}_{21}) + 2\hat{\pi}_{12}\hat{\pi}_{21}]}$$

where $\hat{\pi}_{ij} = x_{ij}/n$. With k independent samples, a linear combination or contrast of differences $L = a_1\Delta_1 + \cdots + a_k\Delta_k$ may be estimated by $\hat{L} = a_1\hat{\Delta}_1 + \cdots + a_k\hat{\Delta}_k$ with standard error

$$\hat{V}^{1/2}(\hat{L}) = \sqrt{a_1^2\,\hat{V}(\hat{\Delta}_1) + \cdots + a_k^2\,\hat{V}(\hat{\Delta}_k)}$$

which is useful for comparing effects across groups.

When some of the subjects have missing responses at time 1 or time 2, the simple analyses described above are no longer straightforward. A complete-case analysis which discards the responses of those missing at either occasion results in loss of power and potential bias. Marascuilo *et al*.[27] analyse the data in Table 1 using a technique of Ekbohm[29] which relies a method-of-moments estimate for Δ. The Ekbohm technique is easy to carry out with hand-calculator operations, but it requires that the missing data be missing completely at random, which is the same assumption that underlies a complete-case analysis. Using software for multiple imputation, one can easily perform a repeated-imputation analysis which is valid under the less restrictive assumption of ignorability.

In this example, the complete data may be expressed as a classification of subjects by four variables: treatment group T (E = experimental, C = control), grade G (6 = sixth, 8 = eighth), response before intervention B (disagree, agree), and response after intervention A (disagree, agree). Using S-PLUS Version 4 and functions from the `CAT` library, it is a simple matter to generate MIs under the general or saturated

multinomial model for T, G, B, A. The `CAT` library implements MI for multivariate categorical data using MCMC techniques described by Schafer[4] (chapters 7–8). I created $m = 5$ imputations by data augmentation runs of 100 cycles each using a standard noninformative Jeffreys prior. Each run was started at the maximum-likelihood parameter estimates obtained from an EM algorithm. The entire process, which consisted of a single run of the function `em.cat` followed by five runs of `da.cat` and `imp.cat`, took approximately 15 s on a 266 MHz Pentium computer. Results of the imputation procedure are shown in Table 2. The `CAT` library is available at `http://www.stat.psu.edu/~jls/misoftwa.html`.

Let us define Δ to be the difference in probability of disagreement after versus before intervention, so that $\Delta > 0$ is the desired outcome. Repeated-imputation estimates for Δ, along with standard errors, degrees of freedom ν, and p-values for testing $\Delta = 0$ against the two-sided alternative are shown in Table 3 for the four $T \times G$ groups. This table also reports the estimated fraction of missing information λ for each estimand as defined in Section 2. The estimated Δ-effects are positive and highly significant in the experimental groups but nonsignificant in the control groups, providing evidence that the experimental treatment was effective. The degrees of freedom ν are all rather large, indicating that the Student's t-approximation (2.4) is

Table 2 Five imputations of the complete data cross-classified by T = treatment group, G = grade, B = response before intervention, and A = response after intervention

				Imputed frequency				
T	G	B	A	1	2	3	4	5
E	6	Disagree	Disagree	68	72	68	71	71
E	6	Disagree	Agree	26	24	25	24	21
E	6	Agree	Disagree	74	64	69	66	67
E	6	Agree	Agree	86	94	92	93	95
E	8	Disagree	Disagree	120	113	120	116	117
E	8	Disagree	Agree	16	16	15	16	16
E	8	Agree	Disagree	44	50	45	49	48
E	8	Agree	Agree	26	27	26	25	25
C	6	Disagree	Disagree	98	97	94	94	95
C	6	Disagree	Agree	24	22	27	19	23
C	6	Agree	Disagree	15	16	17	18	16
C	6	Agree	Agree	23	25	22	29	26
C	8	Disagree	Disagree	135	140	139	135	136
C	8	Disagree	Agree	46	40	39	37	44
C	8	Agree	Disagree	32	31	34	37	35
C	8	Agree	Agree	49	51	50	53	47

Table 3 Repeated-imputation inferences for Δ-effects by treatment group and grade

	Estimate	SE	ν	p	λ
Δ_1: experimental, grade 6	0.173	0.039	258	0.00	0.13
Δ_2: experimental, grade 8	0.152	0.039	350	0.00	0.11
Δ_3: control, grade 6	−0.041	0.046	53	0.37	0.30
Δ_4: control, grade 8	−0.028	0.040	43	0.48	0.34

Table 4 Repeated-imputation inferences for some contrasts of interest

Contrast	a_1	a_2	a_3	a_4	Estimate	SE	ν	p	λ
Experimental (E) vs control (C)	1	1	−1	−1	0.395	0.084	63	0.00	0.28
E vs C, grade 6	1	0	−1	0	0.214	0.063	50	0.00	0.31
E vs C, grade 8	0	1	0	−1	0.181	0.053	252	0.00	0.13
grade 6 vs grade 8	1	−1	1	−1	0.008	0.078	206	0.92	0.15
grade 6 vs grade 8, E group	1	−1	0	0	0.021	0.058	105	0.72	0.21
grade 6 vs grade 8, C group	0	0	1	−1	−0.013	0.054	279	0.81	0.13
Treatment group × grade	1	−1	−1	1	0.034	0.080	125	0.67	0.19

essentially a normal one. Note that these degrees of freedom, unlike those in linear regression and analysis of variance, do not depend on the sample size but on the number of imputations m and the ratio of between-imputation variance B to within-imputation variance $\bar{U}$.

Inferences for some contrasts $L = a_1\Delta_1 + a_2\Delta_2 + a_3\Delta_3 + a_4\Delta_4$ of interest are shown in Table 4. The first three contrasts are highly significant, indicating that the experimental group performs better than the controls for both grades combined and for each grade individually. The nonsignificance of the last four contrasts indicate no evidence of any differences due to grade. These results obtained through MI agree rather closely with those reported previously.[27] Note that because the imputation model is quite general, the imputations in Table 2 may be used for a variety of other analyses as well, such as logit modelling of the post-test response.

Acknowledgements

This work was supported by grant 1-P50-DA10075-01 from the National Institute on Drug Abuse and grant 2R44CA65147-02 from the National Cancer Institute. Thanks to Xiao-Li Meng and John Barnard for helpful comments.

References

1 Madow WG, Nisselson H, Olkin I eds. *Incomplete data in sample surveys, Vol 1: report and case studies*. New York: Academic Press, 1983.

2 Little RJA, Rubin DB. *Statistical analysis with missing data*. New York: John Wiley, 1987.

3 Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley, 1987.

4 Schafer JL. *Analysis of incomplete multivariate data*. London: Chapman & Hall, 1997.

5 Little RJA. Survey nonresponse adjustments. *International Statistical Review* 1986; **54**: 139–57.

6 Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**: 846–66.

7 Dempster AP, Laird N, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 1977; **39**: 1–38.

8 Barnard J, Rubin DB. Small sample degrees of freedom with mutiple imputation. *Biometrika* 1999, in press.

9 Li KH, Raghunathan TE, Rubin DB. Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association* 1991; **86**: 1065–73.

10 Li KH, Meng XL, Raghunathan TE, Rubin DB. Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica* 1991; **1**: 65–92.

11 Meng XL, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* 1992; **79**: 103–11.
12 Fay RE. When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1992: 227–32.
13 Meng XL. Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* 1995; **10**: 538–73.
14 Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**: 473–89.
15 Graham JW, Schafer JL. On the performance of multiple imputation for multivariate data with small sample size. In: Hoyle R ed. *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage, 1998.
16 Schafer JL, Schenker N. Inference with imputed conditional means. *Journal of the American Statistical Association* 1999; **94**: in press.
17 Gilks WR, Richardson S, Spiegehalter DJ eds. *Markov chain Monte Carlo in practice*. London: Chapman & Hall, 1996.
18 Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* 1987; **82**: 528–50.
19 Schafer, JL. Imputation of missing covariates under a multivariate linear mixed model. Technical report 97–10, Methodology Center, Penn State University, available at `http://methcenter.psu.edu`.
20 Data Analysis Products Division, Mathsoft Inc. *S-PLUS user's guide* 1977. Seattle, WA: Mathsoft.
21 SPSS Inc. *SPSS missing value analysis 7.5*. Chicago, IL: SPSS.
22 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
23 Lavori PW, Dawson R, Shera D. A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine* 1995; **14**: 1913–25.
24 Statistical Solutions, Inc. *Solas for missing data analysis Version 1*. Cork: Statistical Solutions, 1988.
25 Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis* 1996; **22**: 425–46.
26 Stata Corp. *Stata user's guide*. College Station, TX: Stata Press, 1997.
27 Marascuilo LA, Omelich CL, Gokhale DV. Planned and post hoc methods for multiple-sample McNemar (1947) tests with missing data. *Psychological Bulletin* 1988; **103**: 238–45.
28 McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **12**: 153–57.
29 Ekbohm G. On testing the equality of proportions in the paired case with incomplete data. *Psychometrika* 1982; **47**: 115–18.