Statistical Methods in Medical Research

http://smm.sagepub.com

Applications of multiple imputation in medical studies: from AIDS to NHANES

John Barnard and Xiao-Li Meng

Stat Methods Med Res 1999; 8; 17
DOI: 10.1177/096228029900800103

The online version of this article can be found at:
http://smm.sagepub.com/cgi/content/abstract/8/1/17

Published by:
SAGE
http://www.sagepublications.com

Statistical Methods in Medical Research 1999; **8**: 17–36

Applications of multiple imputation in medical studies: from AIDS to NHANES

John Barnard Department of Statistics, Harvard University, Massachusetts, USA and
Xiao-Li Meng Department of Statistics, The University of Chicago, Illinois, USA

Rubin's multiple imputation is a three-step method for handling complex missing data, or more generally, incomplete-data problems, which arise frequently in medical studies. At the first step, $m(>1)$ completed-data sets are created by imputing the unobserved data m times using m independent draws from an imputation model, which is constructed to reasonably approximate the true distributional relationship between the unobserved data and the available information, and thus reduce potentially very serious nonresponse bias due to systematic difference between the observed data and the unobserved ones. At the second step, m complete-data analyses are performed by treating each completed-data set as a real complete-data set, and thus standard complete-data procedures and software can be utilized directly. At the third step, the results from the m complete-data analyses are combined in a simple, appropriate way to obtain the so-called repeated-imputation inference, which properly takes into account the uncertainty in the imputed values. This paper reviews three applications of Rubin's method that are directly relevant for medical studies. The first is about estimating the reporting delay in acquired immune deficiency syndrome (AIDS) surveillance systems for the purpose of estimating survival time after AIDS diagnosis. The second focuses on the issue of missing data and noncompliance in randomized experiments, where a school choice experiment is used as an illustration. The third looks at handling nonresponse in United States National Health and Nutrition Examination Surveys (NHANES). The emphasis of our review is on the building of imputation models (i.e. the first step), which is the most fundamental aspect of the method.

1 Introduction and overview

1.1 Difficulties with incomplete-data problems

Missing data or more generally incomplete data (e.g. censored data which are not completely missing because we know which intervals they fall into) occur frequently in medical studies, in the forms of nonresponse in patient surveys, noncompliance in clinical trials, nonreporting or delayed reporting to health surveillance systems, just to list a few. Three major difficulties with such incomplete-data problems are: (I) loss of information, efficiency or power due to loss of data; (II) complication in data handling, computation and analysis due to irregularities in the data patterns and non-applicability of standard software; and most fundamentally; and (III) potentially very serious bias due to systematic differences between the observed data and the unobserved data.

The best way of dealing with these problems, of course, is to avoid them in the first place. Unfortunately, in most real-life studies, be they medical or otherwise, the problem of incomplete data is unavoidable even if we have made the greatest possible efforts. For example, it is a common knowledge that no real-life questionnaire survey can achieve a response rate that is remotely close to 100%. In fact, the problem is so

Address for correspondence: Xiao-Li Meng, Department of Statistics, The University of Chicago, IL 60637, USA. E-mail: meng@galton.uchicago.edu

 0962-2802(99)SM173RA

universal that an unusually high response rate (e.g. 95%) should make the investigator worry about possible design flaws in the survey, such as selection bias in the sample or a substantial amount of untrustworthy responses induced by too much monetary incentive. As another example, even very carefully designed and implemented clinical trials often face problems such as censored data, attrition, and noncompliance. These are all examples of incomplete-data problems, a term we use in its broadest sense. That is, it even includes problems of missing observations that are induced by a study design or are inherently unobservable (e.g. the 'would-be' response to a treatment for those who are assigned to the control group; see Section 3).

Once the data collection process (including follow-up) is completed, there is little one can do about problem (I) besides using the experience to create better designs for future similar studies. Problem (II) is the most visible one in practice, and thus has received the most complaints and most research attention. Problem (III) is most fundamental because if the observed data represent a biased sample of what we intend to study, then without efforts to reduce such bias, our inference would not be of much scientific value regardless of the sophistication of our computational method or analysis procedure. Unfortunately, it is also the most difficult one to handle because, typically, the reasons for not observing the full data (i.e. the so-called missing-data mechanism) are often at best partially understood (except for cases where missing data are induced by the design or latent-variable modelling). It is also certainly one of the most overlooked problems in practice. It is not uncommon for an investigator to analyse whatever is available (as with the 'complete-case' methods and the 'available-case' methods) and never even realize the potentially serious nonresponse bias.[1]

1.2 Rubin's multiple imputation method

Imputation, that is, filling in missing data by some plausible values, has been a popular method for handling incomplete-data problems. This popularity largely stems from the fact that once the missing values are filled in, standard complete-data methods can be readily applied to produce 'results', and thus problem (II) is avoided. However, in order to have an inferentially useful analysis based on data sets that are partially imputed, two requirements must be met. First, the imputation method/model must reasonably capture the actual distributional relationships between the unobserved and the observed. Secondly, the analysis must take into account the uncertainty in the imputed values, because no matter how much effort one makes, the imputed values are simply not the real observations.

Rubin's multiple imputation[2,3] is a three-step method for meeting these two requirements that maintains the principal attraction of an imputation method. The first step is to build a *sensible* imputation model – the meaning of 'sensible' will be discussed in Section 1.3 – and then to impute the unobserved values by $m(>1)$ independent draws from the model. We thus have m completed data sets, $Y^{(\ell)} = \{Y_{mis}^{(\ell)}, Y_{obs}\}$, $\ell = 1, \ldots, m$, where Y_{obs} is the observed data, and $Y_{mis}^{(\ell)}$ is the ℓth imputation of Y_{mis}, the missing data.

At the second step, we simply conduct a complete-data analysis using each $Y^{(\ell)}$ in the same way as we would use a real complete data set Y. Let θ be a d-dimensional unknown quantity of interest (e.g. a regression slope, β), $\hat{\theta}(Y)$ be an efficient estimate of θ based on Y (e.g. $\hat{\beta}$), and $U(Y)$ be the associated variance (e.g. an estimate of

$\mathrm{Var}(\hat{\beta})$). Then our m completed-data analyses produce

$$\hat{\theta}_\ell \equiv \hat{\theta}(Y^{(\ell)}) \quad \text{and} \quad U_\ell \equiv U(Y^{(\ell)}), \quad \ell = 1, \ldots, m$$

At the third step we combine these quantities to obtain the so-called repeated-imputation inference, as defined by Rubin.[3] The combining rule is most straightforward for the point estimate of θ, which is simply the average of $\{\hat{\theta}_\ell, \ell = 1, \ldots, m\}$

$$\bar{\theta}_m = \frac{1}{m} \sum_{\ell=1}^{m} \hat{\theta}_\ell$$

The combining rule is also straightforward for the variance estimate for $\bar{\theta}_m$

$$T_m = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m$$

where $\bar{U}_m$, the average of $\{U_\ell, \ell = 1, \ldots, m\}$, estimates the *within-imputation* variability

$$B_m = \frac{1}{m} \sum_{\ell=1}^{m} (\hat{\theta}_\ell - \bar{\theta}_m)(\hat{\theta}_\ell - \bar{\theta}_m)^\top$$

estimates the *between-imputation* variability, and the inflation factor $(1 + m^{-1})$ accounts for the additional variability due to using a finite number of imputations (in contrast to using an infinite number of imputations).

Methods for constructing confidence intervals and for computing p-values are slightly more complicated because one cannot simply use a normal approximation based on $\bar{\theta}_m$ and T_m, which would be appropriate if both n and m were large, where n is a measure of the size of Y_{obs}. In practice, m is often small to moderate (e.g. between five and ten), and thus we need to take into account the 'degrees of freedom'. Several simple rules/methods are available for dealing with different settings. They include the large-n rule of Li *et al.*,[4] the small-n rule of Barnard and Rubin,[5] and the likelihood-ratio method of Meng and Rubin.[6] See Schafer's article in this issue and Rubin and Schenker[7] for a review of these methods. Rubin and Schenker[7] also contains an overview of some applications of the multiple imputation approach in health care research up to 1991.

Thus, Rubin's multiple imputation method meets the requirement of properly accounting for uncertainty in the imputed values through the use of the easily computable between-imputation variability measure B_m, which is not available with single imputation. It is also clear that the method maintains and in fact enhances the simplicity of single-imputation methods because at the analysis stage (i.e. the second and third steps), no special incomplete-data procedures are needed.

1.3 Building a sensible imputation model

The requirement for constructing a sensible imputation model that reasonably depicts the underlying missing-data mechanism is a common requirement for any useful imputation method. By 'sensible' we mean a model that incorporates as much as possible the available data and our knowledge about the missing-data mechanism, but

at the same time keeps the model building and fitting feasible. This balance is important in practice because an overly simplistic model, such as imputing missing values by some sort of average of observed data, is typically completely inadequate for imputation. For one thing, these averages are subject to the same nonresponse bias we intend to reduce. On the other hand, an overly complex model, such as a model involving many high order interactions, will not only face the problem of nonidentifiability and other difficulties in model building, but more importantly, an overly complex model could have poor prediction power because it over-fits the existing data (these problems can be avoided when using a full Bayesian analysis and various model diagnostics; see Section 5). Furthermore, over-complexity often makes a method less acceptable in practice and more vulnerable to implementation errors.

While it is impossible to have a general recipe for achieving this balance, there are some general guidelines that have been emphasized in the literature.[3,8,9] In the context of producing imputations for public-use data files (e.g. the application in Section 4), the imputation model should satisfy what Meng[9] called the *practical objectivity and generality* requirement, meaning that the model should not be in serious conflict with common analytic models used for analysing the data files. For example, the model should not restrict a two-way interaction to be zero when that very interaction is of common interest to analysts. In other words, the imputer should make efforts to accommodate common models used in analyses, even if some of these are not sensible for the purpose of imputation – Meng[9] gives a general discussion of these and related issues.

On the other hand, if the imputer is the same party as the analyst with a specific analysis goal (e.g. the application in Section 2), then the imputation model can be more tailored. Whereas Rubin's multiple imputation method was originally motivated by the need for handling nonresponse in public-use data files or shared databases, there has been increasing use of the method for more small scale studies, including traditional 'one-analyst-one-goal' studies. In such studies, Rubin's method offers attractive flexibility in both model building and computation via the separation of the task of handling incomplete data from the task of answering the questions of substantive interest. This is so called 'in-house' use of multiple imputation, a term that is used to distinguish it from the original applications for creating imputations for general 'outside' users.

With an 'in-house' application, among the class of models that are reasonable for imputation, the analyst should choose the one that is easiest to implement and is in as little conflict as possible with his or her particular analysis model. Note that the two models can be, and often are, incompatible with each other because a model useful for imputation may not be a model of substantive interest and vice versa. This is the issue of *uncongeniality* discussed in detail by Meng,[9] whose theoretical results show that uncongeniality typically leads to conservative inference, assuming, of course, both models are reasonably constructed. Rubin[8] gives related discussions, especially regarding the issue of distinguishing between achievable objectives and ideal but unrealistic objectives in the context of handling incomplete data.

These points will be illustrated in detail in the next three sections. In Section 2, we demonstrate how the problem of reporting delay in estimating the survival time after AIDS diagnosis can be handled by Rubin's method; this is an example of 'in-house'

use. In Section 3, we discuss the use of Rubin's method as an integrated part of a general methodology for analysing randomized experiment data in the presence of both noncompliance and missing observations. This is an example of 'in-between' use, for the method is neither used for a specific 'in-house' application nor is it used for generating imputations for 'outside' users of a public-use database, but rather as a part of a general methodology for a class of problems. In Section 4, we demonstrate a classic 'outside' application of Rubin's method for multiply imputing the nonresponse in the US National Health and Nutrition Examination Surveys (NHANES).

2 Application I: handling reporting delay for AIDS survival estimation with surveillance data

2.1 Difficulties with surveillance data

Data collected and maintained by national and regional health surveillance systems are the basis for estimating prevalence, mortality rate, survival times and other vital features of various diseases occurring in general populations, which in turn are the basis for assessing general health care needs, long-term health-policy planning, general disease-prevention education, etc. For example, the data collected and maintained by the AIDS surveillance system of the US Centers for Disease Control (CDC) made it possible to obtain estimates of the mortality and survival times of the general AIDS population residing in the United States, estimates that are of vital importance for studying the AIDS epidemic in the United States. However, estimation with such surveillance data is a very challenging task because of severe incompleteness of the data in various aspects.

For example, because a non-negligible fraction of deaths among the reported AIDS cases to CDC will never be reported to CDC, survival-time estimates based on all reported AIDS cases without death certificates censored at the time of analysis are biased upward. An example of this sort is given by Tu *et al.*,[10] who show that such a method would estimate the five-year (after diagnosis) survival rate for *Pneumocystis carinii* pneumonia (PCP), an AIDS-defining disease, to be over 15%, which was clearly an overestimation when compared to the reality. Such an overestimation contributed to some early unfounded optimism about the AIDS epidemic and treatments.[11]

To overcome this bias problem, Tu *et al.*[10,12] adopted the strategy of using only the reported deaths, and thus the never-reported death cases were excluded. While this approach itself is subject to bias if those never-reported cases have systematically different survival times from those that were reported, which is entirely possible, the results obtained by Tu *et al.*[10,12] were much closer to reality judging by their comparability with findings in several clinical trials.[11,13]

However, the reported-death approach also faced several challenging incomplete-data problems, most crucially the problem of delay in reporting death to CDC. The completely observed cases were those who were diagnosed with AIDS, died, and were reported to CDC by a chronological time t^* defined by the analysis. If our analysis only uses these cases, then our results will be contaminated with biases of various sorts, such as likely underestimation from those who were diagnosed and died earlier in the

epidemic, and possible overestimation from those who resided at geographic locations with better health surveillance systems; these biases are impossible to disentangle and remove without additional information and analysis.

Ideally, the reporting delay can be handled by jointly modelling and estimating the survival and reporting delay distributions given the time of an AIDS diagnosis, the time of death, and the reporting time of death. Unfortunately, this joint estimation is difficult to implement because the reporting time of death is not available for some of the reported deaths. At the time of the analysis conducted by Tu *et al.*,[10,12] the reporting time was known only for deaths occurring after September 1987, while the data set available for analysis consisted of 82 239 reported deaths for the male population and of 6566 for the female population between the first quarter of 1983 and the first quarter of 1991 (paediatric and transfusion-related AIDS cases were excluded). A further complication is that any death occurring before 1984 has an unknown time of death, though this is a relatively minor complication because those cases, being relatively few, can be discarded without appreciable effect on the general conclusions.

2.2 A multiple-imputation strategy and model assumptions

Rubin's multiple imputation method offers an attractive alternative for handling such reporting delay problems. As detailed by Tu *et al.*,[10,12] this approach allows us to separately deal with the reporting delay and with the survival analysis, thus, avoiding the difficulties of joint estimation. We can first concentrate on the modelling of the survival time without worrying about the reporting delay, which obviously is a nuisance to our main objective. We then estimate the reporting delay distribution using available information, multiply impute the delayed cases, and proceed with the second and third steps of Rubin's method to obtain our inference about the survival time. Specifically, our strategy includes the following five steps:

Step 1 construct a survival-time model pretending there were no reporting delays;
Step 2 construct and estimate the reporting-lag model using available information;
Step 3 multiply impute the delayed cases using the imputation model built upon Step 2;
Step 4 compute the estimates of the parameters of the survival-time model as well as their variance estimates using each of the completed-data sets;
Step 5 use the repeated-imputation inference rules to combine the estimates found in step 4 to compute our final estimates of the model parameters and their variances, and then draw inferences from the fitted survival-time model.

For both steps 1 and 2, because the times (e.g. death time; reporting time) were recorded in quarters, Tu *et al.*[10] adopted a discrete-time proportional hazards model of the following form:

$$\text{Prob}(T = j \mid z, \theta) = \begin{cases} (p_0 \cdots p_{j-1})^{\exp\{z^\top \beta\}} (1 - p_j^{\exp\{z^\top \beta\}}) & \text{if } 0 \le j \le \mathcal{J} - 1 \\ (p_0 \cdots p_{\mathcal{J}-1})^{\exp\{z^\top \beta\}} & \text{if } j = \mathcal{J} \end{cases} \tag{2.1}$$

In equation (2.1), T is a time variable (e.g. the survival time after AIDS diagnosis S, or the lag time of death reporting R), z is a vector of covariates, $\mathcal{J}$ is the maximum

observed number of quarters for T, and $p_j = \text{Prob}(T > j+1 \mid T > j)$ $(0 \leq j \leq \mathcal{J}-1)$ are conditional baseline probabilities. The model parameter θ is parameterized as $\theta = (\alpha, \beta)$ with $\alpha = (a_0, \ldots, a_{\mathcal{J}-1})$ and $a_j = \log[-\log(p_j)]$ for $1 \leq j \leq \mathcal{J}-1$. This reparametrization of $p = (p_1, \ldots, p_{\mathcal{J}-1})$ not only facilitates computation by removing the range restriction on p (i.e. $0 \leq p_j \leq 1$), but more importantly, allows for a better normal approximation to the likelihood function, or more generally, the posterior distribution of θ. The normal approximation to the likelihood is a common approximation that justifies the use of maximum likelihood estimates and simplifies our imputation procedure (see Section 2.3).

The proportional hazards model is a very common model for survival analysis,[14,15] and the discrete analogue (2.1) was suitable for the current problem because of the discrete nature of the data. The model is quite suitable for comparing survival times among various subpopulations defined by factors such as risk groups and geographical location via the use of covariate z, especially because it fits these subpopulation data reasonably well, as discussed by Tu *et al.*[12] The model is also used for modelling the reporting-lag time mainly for two reasons. The obvious reason is that this greatly simplifies the computation because only one subroutine is needed for fitting both models. The more important reason is that the reporting-lag time is just a type of survival time if we consider 'not reporting' equivalent to 'survive'. For example, it makes good sense to consider the hazard function for not reporting, that is, the probability that a death will be reported at quarter $j+1$ given it has not been reported up to quarter j. The flexibility in the choice of z also allows better estimation of the reporting-lag distribution by accommodating important factors such as geographical location and risk group.

For the male population, Tu *et al.*[10] considered four major risk groups based on sexual behaviour and injecting drug (ID) use status: heterosexual versus homosexual/bisexual crossed with the ID users versus nonusers. For the female population, Tu *et al.*[12] considered two risk groups: the ID users and nonusers. Within each of these groups, three diagnosis strata were considered: PCP, Kaposi's sarcoma, and others. Geographical location was classified into six regions: north-east, central, west, south, mid-Atlantic, and a residual category that consists of areas with population less than one million. Time was also an important covariate – for the survival-time model, the time of diagnosis, and for the reporting-lag model, the time of death. Also considered was an indicator for the change in the definition of AIDS that took place in 1987.

2.3 Model fitting and creating multiple imputations

Having chosen the model including covariates, the popular EM algorithm[16] was used to fit both the survival-time model and the reporting-lag model. (Meng[17] presents a historical link between the EM algorithm and medical studies as well as an overview of EM with various references.) The EM algorithm was chosen for its ability to deal with the problems of right truncation and of left censoring in the data. The right truncation occurred because we can only observe cases whose diagnosis time, death time, and death reporting time were all before the analysis time, t^* (the issue of identifying underlying distributions beyond the truncation point is discussed in Tu *et al.*[10]). The left censoring occurred because of the unknown death time for those who died before 1984, and more importantly, for estimating the reporting-lag distribution,

because of the inclusion of the deaths occurring after January 1986 but before October 1987. The reporting-lag times of these deaths were only known to be less than the time difference between the time of death and the first quarter of 1991 (thus they are left censored), and these cases were included so that estimation of the reporting delay distribution up to five years was possible. The five-year upper bound was used after careful examination of the CDC data, which indicated that deaths not reported within five years would very likely never be reported.

To create multiple imputations for the delayed cases, we first note that if we know the reporting-lag probability $\text{Prob}(R = r \mid z, \theta)$ for any r, then a natural estimate of the expected number of unreported cases with death time t_i and covariate z_i given n_i observed deaths with the same death time and covariate value is given by

$$E[k \mid n_i, P(t_i, z_i, \theta)] = \frac{1 - P(t_i, z_i, \theta)}{P(t_i, z_i, \theta)} n_i \tag{2.2}$$

where k denotes the number of unreported deaths and

$$P(t, z, \theta) = \text{Prob}(R \leq t^* - t \mid z, \theta) = \sum_{r=0}^{t^*-t} \text{Prob}(R = r \mid z, \theta)$$

That is, the ratio of the expected number of unreported cases to the observed cases should be equal to the odds of not reporting by time t^*. A natural model for k that is consistent with (2.2) and common for modelling data of this sort is the negative-binomial distribution

$$NB(k \mid n_i, P(t_i, z_i, \theta)) = \binom{k + n_i - 1}{k} [1 - P(t_i, z_i, \theta)]^k P(t_i, z_i, \theta)^{n_i}$$

Since we only have estimates of θ and thus of $\text{Prob}(R = r \mid z, \theta)$, we need to take into account the uncertainty in estimating θ when we create imputations. This can be achieved by the Bayesian method by drawing θ from its posterior distribution given the available information. This can be done in several ways, including the powerful Markov chain Monte Carlo method.[18] Tu *et al.*[10,12] adopted the traditional large-sample method by using a multivariate normal approximation to the desired posterior of θ, which was adequate given the large size of the data. Specifically, they approximated the posterior distribution by $N(\widehat{\theta}, \widehat{\Omega})$, where $\widehat{\theta}$ and $\widehat{\Omega}$ are, respectively, the maximum likelihood estimate of θ and the inverse of the observed Fisher information matrix obtained in step 2 of Section 2.2.

With these build-ups, we can obtain m completed-data sets by performing the following three steps m times independently. Namely, for each ℓ $(1 \leq \ell \leq m)$:

Step 1 draw a random sample θ_ℓ from $N(\widehat{\theta}, \widehat{\Omega})$;

Step 2 given θ_ℓ, for each observed n_i with death time t_i and covariate z_i, draw a random sample $k_i^{(\ell)}$ from the negative-binomial distribution $NB[k \mid n_i, P(t_i, z_i, \theta_\ell)]$;

Step 3 for each observed n_i with death time t_i and covariate z_i, impute the unreported number of death by $k_i^{(\ell)}$ to form the ℓth completed-data set $Y^{(\ell)} = \{n_i + k_i^{(\ell)}, i = 1, ...\}$.

Since the maximal reporting lag is about five years, only the unreported deaths with an AIDS diagnosis after January 1986 need to be imputed, in which case exact death times (in quarters) are available for the reported ones.

The number of multiple imputations, m, usually does not need to be large to obtain stable answers, though, obviously, the larger the better as long as it is computationally affordable. Tu *et al.*[10] found that results with $m = 10$ and $m = 50$ were similar. It has been a common experience that with many practical problems $m = 5 - 10$ is adequate. The final results of Tu *et al.*[10,12] were based on $m = 50$ and were quite different from using mean imputation, i.e. imputing k by the estimated expected value given by (2.2). Tu *et al.*[10,12] give detailed discussions of this comparison as well as many other results.

3 Application II: handling noncompliance and missing data in a randomized experiment

3.1 Introduction

One of the most pervasive tasks in medical research is trying to draw causal inferences. An ideal scenario for obtaining valid causal inferences for a binary treatment is the following: (1) the data arise from a randomized experiment with two treatments; (2) the outcome variables are fully observed; (3) there is full compliance with the assigned treatment; (4) the design variables are fully observed; and (5) the background variables are fully observed. Aspect (5) is useful for doing covariate adjustment and subpopulation analyses. For this ideal scenario, there are standard and relatively simple methods for obtaining valid causal inferences. In reality, however, particularly with human subject trials, this scenario rarely occurs.

Deviations from the ideal scenario that occur frequently in medical experiments are the following: (a) there exist missing values in the outcomes; (b) there exist missing values in the background variables; and (c) there is noncompliance with assigned treatment. Handling these complications in a valid and general manner is challenging. Here we review a multiple-imputation based framework, proposed by Barnard *et al.*[19] for analysing a randomized experiment suffering from (a), (b), and a special form of (c).

The motivating application in Barnard *et al.*[19] is the Milwaukee Parental Choice Program (MPCP), a randomized experiment in which one of the goals is to estimate the causal effects of 'school choice', e.g. the effects on achievement tests of attending a private school versus attending a public school. The MPCP, while not a medical study, has many of the same features and complications of clinical trials. We use the MPCP to illustrate some of the modelling and computational issues surrounding the use of multiple imputation for handling difficulties (a), (b), and (c) in randomized experiments.

3.2 The Milwaukee Parental Choice Program

The MPCP was launched in 1990 as the first publicly funded school voucher program in the United States. Eligible students from the Milwaukee public school (MPS) system were given the opportunity to attend one of several participating private

schools in the area, labelled 'Choice' schools. Eligibility into the programme was determined predominantly by income level; specifically, household income could not exceed 1.75 times the poverty line. A lottery was held to determine admittance into Choice schools because only a limited number of slots were available in each school and grade per year. The programme was financed by diverting government funding for all accepted students from the public schools they would have attended to the private schools to which they were admitted.

The lottery structure of the admittance procedure represents a naturally occurring randomized block design (in contrast to a planned design). Two of the blocking variables, grade and year of application, were recorded for all participants. Because the school attended by each Choice student was not released by the principal investigator of the study due to confidentiality concerns, Barnard *et al.*[19] and other analysts of the MPCP used race as a proxy for the third blocking variable, school. The use of race as a proxy for school is a reasonable approximation because the three Choice schools that captured over 80% of the Choice students in 1990 were each predominantly either black or Hispanic.[20] Barnard *et al.*[19] restricted their examination of the MPCP to a subsample of experimental subjects, black and Hispanic students from kindergarten through eighth grade. Their subsample consisted of 1151 subjects, with 238 in the control group (MPS schools) and 913 in the treatment group (Choice schools).

The outcome variables of the experiment were normalized scores from standardized math and reading tests, the Iowa Test of Basic Skills (ITBS), taken every spring (beginning after entry into the study). Because funding for the evaluation was provided only through 1994, the maximum number of years of outcome data was four, which was only possible to observe for those who applied in the first year of the program, 1990. Covariate data were also collected on background characteristics of participating students through surveys and administrative records.

Figure 1 describes the variables used by Barnard *et al.*[19] along with their levels of missingness. Many of the covariates and outcomes had rather high levels of missingness, much higher than in a typical medical study, and there were over 100 missing data patterns. Unlike in many longitudinal studies, the missing data pattern of the outcomes variables was far from monotone, which complicates the generation of multiple imputations of the missing data.[21]

In summary, the MPCP can be viewed as a longitudinal study with multiple outcomes, two treatments, and a randomized block design that suffers from incomplete outcomes with a nonmonotone missing data pattern and noncompliance with assigned treatment.

3.3 Noncompliance and potential outcomes

Handling noncompliance in randomized experiments has recently become a topic of great interest.[22–30] A major impetus for this interest is the desire to estimate average causal effects for the complier subpopulation, which is often of greater scientific interest than the usual intention-to-treat estimands.[19] Here we examine the implications of noncompliance for the MPCP.

In their analysis of the MPCP, Barnard *et al.*[19] assumed that compliance, denoted by C, can be viewed both as a characteristic of an individual and expressed as a single binary variable, even though compliance behaviour can vary over time. In addition

Group	Group Notation	% Missing	Variable Description
	Z	0	treatment: MPS (0) or Choice (1)
	C	21	compliance status: never taker (n) or complier (c)
Design Variables	$\underline{W}$	0	race: Hispanic or black
		0	year applied to program: 1990-1993
		0	grade when applied to program: Kg-8
Covariates	$\underline{X}$	0	sex: female or male
		0	lunch: qualifies for free/subsidized
		61	hear: ways people learned about Choice program
		62	employ: 0 vs. ≥ 1 parents employed
		72	relation: parent vs. nonparent guardian
		60	1st language: Spanish or English
		61	religion: Judeo-Christian or other
		60	married: parents married or not
		76	level of father's educational attainment
		60	level of mother's educational attainment
		3	distance: home to closest Choice school
Control Outcomes	$\underline{Y}(0)$	58-96	math test scores after assignment to MPS
		47-90	reading test scores after assignment to MPS
Treatment Outcomes	$\underline{Y}(1)$	42-90	math test scores after assignment to Choice
		46-88	reading test scores after assignment to Choice

Figure 1 Variable missing rates and descriptions of the variables used by Barnard *et al.*[19] in their analysis of the MPCP. The underline indicates that the quantity is a vector. The variables listed in the rows belonging to a group can be collectively denoted by the corresponding symbol in the group notation column. For example, the three design variables are collectively denoted by $\underline{W}$, which is a vector of length three. The vector $\underline{Y}(0)$ and the vector $\underline{Y}(1)$ are both of length eight, because there are potentially two outcomes recorded at four time points. The percentage of missing values for the control and treatment outcomes are ranges over the four years of outcomes.

they assumed that no subjects would take the treatment if assigned to the control group. Under these assumptions, it is possible to classify every individual in the MPCP into one of two compliance categories: compliers ($C = c$) will take treatment if assigned to it (in some average sense over time); never takers ($C = n$) will not take treatment if assigned to it. Further, it is possible to observe the compliance status of individuals assigned to the treatment group (i.e. for students who won the lottery), as the behaviour when assigned to treatment is observed for those in the treatment group; for MPCP, about 30% (284 subjects) of those who were assigned to the treatment group were labeled never takers. The compliance values of individuals assigned to the control group (i.e. for students who lost the lottery), however, are not observed, as it is not known how they would behave when assigned to the treatment group. Hence, questions concerning compliance can be viewed as missing data questions.

To clarify the issues surrounding noncompliance, it is helpful to adopt the potential outcomes notation of 'Rubin's Causal Model'.[31–34] In the current setting of a binary treatment variable, each subject's potential outcomes are given by $(\underline{Y}(0), \underline{Y}(1))$, where

$\underline{Y}(z)$ is the outcome if assigned to treatment $Z = z$, with

$$Z = \begin{cases} 1 & \text{if assigned to treatment} \\ 0 & \text{if assigned to control} \end{cases}$$

For every subject, only one of the two potential outcomes are observed. Hence, each row of the matrix of outcomes $\mathbf{Y}$, consists of observations either of $\underline{Y}(0)$ or of $\underline{Y}(1)$, where the potential outcome contributed by a subject depends on the subject's treatment assignment. In causal studies (e.g. randomized experiments), a common goal is to estimate the average of $\underline{Y}(1) - \underline{Y}(0)$ (regardless of compliance status), usually called the intention-to-treat (ITT) estimand. In casual studies suffering from noncompliance, an additional estimand of interest is the average of $\underline{Y}(1) - \underline{Y}(0)$ for those who are compliers (i.e. the average effect of the treatment for those who take the treatment if assigned to it), called the complier average causal effect (CACE).[23] For example, in the MPCP, an estimand of great interest is the causal effect of attending a Choice school (the treatment) on performance outcomes for those children who would attend a Choice school if asked to attend such a school (i.e. compliers). Note that under the assumption of no defiers (i.e. those who always take the opposite treatment than the assigned) nor always takers, the ITT estimand can be written as a weighted sum of the CACE and the never-taker average causal effect (which is assumed to be zero in Section 3.4).

When some of the outcomes are missing, there is also a pair of corresponding potential nonresponse patterns: $\underline{Ry}(0)$ is the nonresponse pattern for $\underline{Y}(0)$, and $\underline{Ry}(1)$ is the nonresponse pattern for $\underline{Y}(1)$. Hence, each row of the matrix of nonresponse indicators $\mathbf{Ry}$, consists of observations either of $\underline{Ry}(0)$ or of $\underline{Ry}(1)$, where the potential nonresponse pattern contributed by a subject depends on the subject's treatment assignment.

To clarify the interactions among noncompliance, missing values, and the experimental design, Figure 2 indicates what is observed, what is missing but intended to be observed, and what is missing but cannot be observed. Barnard *et al.*[19] impute only the missing intended data and the missing compliance values, i.e. they impute $\mathbf{Y}_{mis}^{\{1c\}}$, $\mathbf{Y}_{mis}^{\{1n\}}$, $\mathbf{Y}_{mis}^{\{0\cdot\}}$, $\mathbf{X}_{mis}^{\{1c\}}$, $\mathbf{X}_{mis}^{\{1n\}}$, $\mathbf{X}_{mis}^{\{0\cdot\}}$, and $\mathbf{C}_{exc}^{\{0\cdot\}}$, but none of the potential outcomes under alternate assignments (i.e. the excluded outcome data). The key point is that given imputations of the intended data (including compliance status), it is then straightforward to get estimates of the ITT estimand and of the CACE.

3.4 Assumptions underlying the multiple imputation approach

In many applications, the first steps in the path to generating multiple imputations are to specify a full model for the complete data and to specify the missing-data mechanism, i.e. the process that generated the missing data. In this section, we review and discuss the complete-data model and missing-data mechanism employed by Barnard *et al.*[19]

In many applications that suffer from missing observations it is generally clear what are the intended complete data. With randomized experiments suffering from noncompliance, however, it is not as clear what is meant by intended complete data. The major question is whether to include compliance status, C, in the list of complete-

Z	C	$\underline{Y}(1)$	$\underline{Y}(0)$	$\underline{R_y}(1)$	$\underline{R_y}(0)$	$\underline{X}$	$\underline{R_x}$	$\underline{W}$
1	**c**	$\mathbf{Y}_{inc}^{\{1c\}} = \begin{bmatrix} \mathbf{Y}_{mis}^{\{1c\}} \\ \mathbf{Y}_{obs}^{\{1c\}} \end{bmatrix}$	$\mathbf{Y}_{exc}^{\{1c\}}$	$\mathbf{R}_{\mathbf{y}_{inc}}^{\{1c\}}$	$\mathbf{R}_{\mathbf{y}_{exc}}^{\{1c\}}$	$\mathbf{X}_{obs}^{\{1c\}}$ $\mathbf{X}_{mis}^{\{1c\}}$	$\mathbf{R}_{\mathbf{x}}^{\{1c\}}$	$\mathbf{W}^{\{1c\}}$
1	**n**	$\mathbf{Y}_{inc}^{\{1n\}} = \begin{bmatrix} \mathbf{Y}_{mis}^{\{1n\}} \\ \mathbf{Y}_{obs}^{\{1n\}} \end{bmatrix}$	$\mathbf{Y}_{exc}^{\{1n\}}$	$\mathbf{R}_{\mathbf{y}_{inc}}^{\{1n\}}$	$\mathbf{R}_{\mathbf{y}_{exc}}^{\{1n\}}$	$\mathbf{X}_{obs}^{\{1n\}}$ $\mathbf{X}_{mis}^{\{1n\}}$	$\mathbf{R}_{\mathbf{x}}^{\{1n\}}$	$\mathbf{W}^{\{1n\}}$
0	$\mathbf{C}_{exc}^{\{0\cdot\}}$	$\mathbf{Y}_{exc}^{\{0\cdot\}}$	$\mathbf{Y}_{inc}^{\{0\cdot\}} = \begin{bmatrix} \mathbf{Y}_{mis}^{\{0\cdot\}} \\ \mathbf{Y}_{obs}^{\{0\cdot\}} \end{bmatrix}$	$\mathbf{R}_{\mathbf{y}_{exc}}^{\{0\cdot\}}$	$\mathbf{R}_{\mathbf{y}_{inc}}^{\{0\cdot\}}$	$\mathbf{X}_{obs}^{\{0\cdot\}}$ $\mathbf{X}_{mis}^{\{0\cdot\}}$	$\mathbf{R}_{\mathbf{x}}^{\{0\cdot\}}$	$\mathbf{W}^{\{0\cdot\}}$

Figure 2 Table of observed, potentially observed, and not possible to observe matrices of values. The rows of the first two columns give the treatment group status and compliance status of the subset of subjects in the corresponding row. The columns headings indicate the type of variable. A bold symbol represents a matrix of values. The subscript *obs* denotes a subset of data actually observed; subscript *mis* denotes a subset of data that is not observed but is possible to observe; subscript *inc* denotes a subset of data that is possible to observe; subscript *exc* denotes subset of data that is impossible to observe. Superscript $\{st\}$ denotes the subset of people with treatment assignment, s (0 = MPS, 1 = Choice), and true compliance type, t (c = complier, n = never taker), with '$\cdot$' indicating the union of subjects over all possible values of the index.

data variables. As noted earlier, if the complete-data estimand is an ITT estimand, knowledge of the compliance values is not necessary, however, it is if the estimand is a CACE estimand. Because of this point, the desired complete data should consist of $\mathbf{Y}_{inc} = [\mathbf{Y}_{inc}^{\{1c\}}, \mathbf{Y}_{inc}^{\{1n\}}, \mathbf{Y}_{inc}^{\{0\cdot\}}]$, the matrix of outcomes that were intended to be collected (see Figure 2) for all subjects, $\mathbf{W}$, the matrix of design variables, $\mathbf{X}$, the matrix of background variables, $\mathbf{Z}$, the matrix of treatment indicators, and $\mathbf{C}$, the matrix of compliance statuses.[19]

The complete-data model used by Barnard *et al.*[19] can be summarized as follows: for each compliance status ($C = c$, complier, or $C = n$, never taker) $f(\underline{X} \mid \underline{W}, C)$ is a general location model, and $f(\underline{Y}(1) \mid \underline{X}, \underline{W}, C)$ and $f(\underline{Y}(0) \mid \underline{X}, \underline{W}, C)$ are multivariate normal distributions, where the parameters of the distributions are implicitly conditioned on and the distributions are allowed to differ across compliance status. For the definition of a general location model, see Section 4.3. They also assume that $f(C \mid \underline{W})$ is a binomial distribution. There is no specification for the the joint distribution of $\underline{Y}(1)$ and $\underline{Y}(0)$, because it is not needed for imputing the missing values in $\mathbf{Y_{inc}}$. Finally, given all of the parameters, each observation is assumed to be independent.

Barnard *et al.*[19] make several critical assumptions about the missing data process. The first is the assumption of 'latent ignorability' of the missing values of covariates and outcomes given true compliance status.[35] That means that if compliance status were fully observed, the missing data mechanism (for Y and X) would be ignorable, that is, the missing data would be missing at random and the parameters of the missing data mechanism would be distinct from the parameters of the data.[1,36] Specifically, they assumed

$$f(\underline{Ry}(1), \underline{Rx} \mid \underline{Y}(1), \underline{X}, \underline{W}, \mathrm{C}) = f(\underline{Ry}(1), \underline{Rx} \mid \underline{W}, C), \text{ and}$$
$$f(\underline{Ry}(0), \underline{Rx} \mid \underline{Y}(0), \underline{X}, \underline{W}, C) = f(\underline{Ry}(0), \underline{Rx} \mid \underline{W}, C)$$

This assumption implies that the missing-data mechanism is not ignorable because C is not observed for the control group, i.e. there is information in the missing data indicators about the missing compliance indicators. While this assumption complicates the generation of multiple imputations of the missing data, it is much more plausible than assuming that the data are missing at random, which is commonly done in randomized experiments with noncompliance.

Their second critical assumption, which involves the missing data mechanism and the complete-data model, is about the effect of treatment assignment on the never takers: the 'compound exclusion restriction',[35] which is an extension of the standard econometric exclusion restriction, as defined by Angrist *et al.*[22] and invoked in several Bayesian analyses.[23,37] The compound exclusion restriction states that for never takers, neither their outcome values nor their missing outcome patterns are affected by treatment assignment. The rationale for this assumption is that because assignment has no effect on treatment received it has no effect on either outcomes or missing data patterns. Specifically, they assumed that

$$\psi(\underline{Y}(1), \underline{Ry}(1) \mid \underline{X}, \underline{Rx}, \underline{W}, C = n) = \psi(\underline{Y}(0), \underline{Ry}(0) \mid \underline{X}, \underline{Rx}, \underline{W}, C = n)$$

where $\psi(A \mid B)$ denote the parameters of the distribution of A given B. The compound exclusion restriction makes it possible to estimate the parameters of $f(\underline{Y}(0) \mid \underline{W}, C = c)$ and of $f(\underline{Ry}(0) \mid \underline{W}, C = c)$ because among the known (observed) compliers there are no observed values of $\underline{Y}(0)$ nor of $\underline{Ry}(0)$ (see Figure 3).

3.5 Generating multiple imputations

Generating multiple imputations under the model outlined in Section 3.4 is challenging. The challenge mainly stems from the two critical assumptions: latent ignorability and the compound exclusion restriction. Imbens and Rubin[23] consider generating imputations for the much simpler case in which the only variable with missing values is compliance and there are no covariates. Barnard *et al.*[19] outline an approach for generating imputations of the missing values for the MPCP, however, their approach turns out to be more complicated than necessary.

Generating imputations of the missing values for the MPCP under the model discussed is most easily accomplished by using Markov chain Monte Carlo methods,[18,21] such as the Gibbs sampler, which can be computationally intensive and often require expert guidance in order to achieve good performance. The implementation details are too involved to describe here, but an outline of a single iteration of a Gibbs sampler for generating imputations is the following:

Step 1 given the compliance indicators, within each compliance status impute the missing values (of outcomes and covariates) and generate a draw of the parameters governing the distributions of the outcomes, covariates, and missing data patterns;

Step 2 given the outcomes, covariates and the distributional parameters from step 1, generate a draw of the missing compliance indicators.

Iterating between these two steps is computationally intensive but relatively straightforward to implement.

Software for performing such simulations for the MPCP and experiments with similar structure is currently under development. Once this software is commonly available, the multiple imputation approach for handling noncompliance and missing data in randomized experiments will provide a powerful and attractive tool for obtaining causal estimates in this difficult but important scenario.

4 Application III: handling nonresponse in NHANES

4.1 Introduction

As we discussed in Section 1, Rubin's multiple imputation method was originally designed to handle nonresponse in surveys that are conducted for constructing public-use data files. Because imputation is generally a model-based procedure while the standard analyses procedures for survey data are typically design-based, there have been a variety of empirical investigations to validate the validity of the repeated-imputation results from the design-based perspective. One such investigation is the simulation study of Ezzati-Rice *et al.*,[38] which used several National Health and Nutrition Examination Surveys (NHANES). The NHANES are periodic surveys conducted by the US National Center for Health Statistics (NCHS) for assessing the health and nutritional status of the US population. Multiple imputation was used to create a CD-ROM research database, to be released soon, for about 70 key survey variables in NHANES-III. However, instead of describing this multiple imputation project, we choose to discuss the simulation study of Ezzati-Rice *et al.*[38] because it illustrates most of the same issues that arose in the process of generating multiple imputations for that database, and it is much easier to describe (e.g. it only involves 17 variables). Here we review some key aspects of this very realistic simulation study, with a particular focus on the survey design and imputation model.

In order to see how repeated-imputation results infer the 'true' values of the estimands, Ezzati-Rice *et al.*[38] constructed a hypothetical population using data from several NHANES to resemble populations surveyed by NCHS. The hypothetical population consisted of 31 847 cases, obtained from four NCHS surveys: HANES-I (1971–74, $N = 11\,678$), HANES-II (1976–80, $N = 10\,371$), NHANES-III, Phase 1 (1988–91, $N = 6874$), and HHANES (Mexican-Americans only; 1982–84, $N = 2944$), where the years give the period over which each survey was collected and N is the number of cases included from that survey. Only adult cases (age greater than 19 years) and those that had no missing values on the ten examination variables (listed below) were included in the population.

The population contained 17 variables in total. Of primary interest are the ten examination variables: standing height, sitting height, weight, systolic blood pressure, diastolic blood pressure, total serum cholesterol, haemoglobin, haematocrit, iron and total iron-binding capacity. The five auxiliary variables, which are fully observed, are described in Figure 3. The two interview variables, self-reported height and self-reported weight, were completely missing in HANES-I and had modest numbers of missing values in the other surveys. These missing values were imputed using a hot-deck procedure (which was used only for the purpose of creating the hypothetical population).

Each simulation of Ezzati-Rice *et al.*[38] consisted of the following four steps:

Variable	Abbreviation	Levels
Age	age	20–39, 40–59, 60–74, 75+
Sex	sex	Female, Male
Race/Ethnicity	race	Black, Mexican American, Other
Location	stdrc	13 unique locations
Household Size	hhsize	1–2, 3–4, 5+

Figure 3 Household screening variables used in the simulation

Step 1 draw a stratified sample from the artificial population using a sample design that mimics the actual designs of NHANES;
Step 2 impose a missing data pattern on the sample from step 1, thus, creating a sample with missing values, using a missing-data mechanism that mimics the missing data patterns observed in the NHANES-III survey;
Step 3 impute five times the missing values in the incomplete sample from step 2, using a general location model;
Step 4 conduct repeated-imputation inference with the five completed data sets created in step 3.

Steps 1–4 were then repeated 1000 times, and the 1000 results from step 4 were compared to the true values from the population (e.g. checking the coverage of repeated-imputation interval estimates). To illustrate how one constructs an imputation model to capture the sample design, the details of steps 1 and 2 are reviewed in Section 4.2, and the imputation model is outlined in Section 4.3.

4.2 Survey design and missing-data mechanism

The survey design of Ezzati-Rice *et al.*[38] was as follows.

1) Classify the 31 847 cases into 48 strata defined by crossing two levels of age (20–59, 60+) with 24 chosen race-location cells.
2) Assign weight to the 31 847 cases in the following manner: unit i in stratum h receives weight

$$w_i^* = N_h \frac{1/\pi_i}{\sum_{j \in h} 1/\pi_j}$$

where N_h is the year 2000 estimated population total for stratum h, and $1/\pi_i$ is the original survey weight for unit i. The weights w_i^* were rounded off to the nearest integer.
3) For $i = 1, \ldots, 31\,847$, include case i in the simulated sample d_i times, where

$$d_i \sim \text{Bin}(w_i^*, n_h/N_h)$$

$\text{Bin}(n,p)$ is a binomial random variable on n trials with success probability p, and n_h is the expected sample size in stratum h for NHANES-97+ (a future NCHS survey at the time of Ezzati-Rice *et al.*[38]). Because the design for NHANES-97+ was not known, n_h was originally approximated by the number of interviewed

persons in stratum h in NHANES-III, Phase 1. Under this plan, the total expected sample size was $\sum_h n_h = 9488$. However, to avoid excessive resampling, the n_h were scaled so that the total expected sample size was 6000.

The missing-data mechanism was constructed in the following way. For each person in the simulated sample, assign the missingness pattern of a randomly selected interviewed person from NHANES-III, Phase 1, in the same response-pattern cell. Response-pattern cells were defined by 576 cells, formed by crossing 24 race-location cells with sex, age , and household size. Cells were then collapsed to yield at least five interviewed NHANES-III persons in each cell.

Note that in the sampling procedure w_i^* is the number of members in the hypothetical population that unit i represents. Hence, we can view the generated samples as coming from a population with $N = \sum_i w_i^* = 196\,478\,806$ members. Note also that the missing data are missing at random but not missing completely at random[36] because the probability of being missing is unrelated to the missing value but depends on the stratification variables.

4.3 Creating multiple imputations under the general location model

The imputation model used by Ezzati-Rice *et al.*[38] was a general location model[21] – the same class of models were used for the actual 1992–93 NHANES-III imputation project.[39] A main reason of using the general location model is that it allows simultaneous modeling of categorical variables and continuous variables, and it is flexible enough for incorporating various constraints for both the categorical variables and continuous variables. For the current application, Ezzati-Rice *et al.*[38] used all 17 variables described in Section 4.1; this is the recommended strategy: use as many variables as possible and feasible.

The five fully observed variables – age, sex, race, stdrc, and hhsize – were treated as categorical in the imputation model. The sample counts were assumed to have a multinomial distribution over the five-way cross-classification of these variables, a frequency table with 936 cells. There were no constraints placed on the cell probabilities besides that they summed to one. Note that the model specification for the frequency table had no impact on the imputations since the five variables involved were fully observed.

The 12 continuous variables were assumed to be conditionally multivariate normal given age, sex, race, stdrc, and hhsize. The conditional means varied across the cells of the frequency table according to a model which included: (a) an intercept; (b) main effects for age, sex, race, stdrc and hhsize; (c) all two-way interactions among age, sex, and race; and (d) the three-way interaction among age, sex and race. The model was the same for the means of all 12 continuous variables. Although the means were allowed to change with the cells of the frequency table, the conditional within-cell covariance matrix (12×12) was assumed to be the same across cells, which is a somewhat undesirable feature of the general location model since it can lead to appreciable under- or over-estimation of variability within individual cells; see Barnard[40] for examples and Liu and Rubin[41] and Barnard *et al.*[42] for relaxations of this assumption. In total, there were 1469 parameters in the multivariate model: 935

cell probabilities, 456 regression coefficients and 78 variances and covariances from the within-cell variance-covariance matrix. Jeffreys' prior distributions were assumed for both the location parameters and the dispersion parameters.

The multiple imputations, five for each simulated dataset, were generated for the above model by drawing from the posterior predictive distribution (i.e. the posterior distribution of the missing values given the observed values). The computations for this drawing were accomplished using the data augmentation technique.[21,38]

The main goal of Ezzati-Rice *et al*.[38] was to demonstrate that the repeated-imputation confidence interval estimates, based on multiple imputations generated under a general location model, had the correct coverage. Their results shown that the imputation model was successful in creating valid design-based repeated-sampling inferences for a wide range of estimands (e.g. population and subdomain means).

5 A cautionary concluding remark

The use of Rubin's method as an 'in-house' analysis tool is a relatively new application of the method; its greatest advantage is its flexibility in separately handling the incomplete-data problems and the substantive analysis. Cautions are needed, however, just as with any statistical methodology. It is clear that if the imputation model is seriously flawed in terms of capturing the missing-data mechanism, then so will be any analysis based on such imputations. This problem can only be avoided by carefully investigating each specific application, by making the best use of knowledge and data about the missing-data mechanism, and by performing various model checking procedures, in particular, posterior predictive checks.[43,44] This is not an additional burden for using Rubin's method, but rather a fundamental requirement for any general method that attempts to produce statistically and scientifically meaningful results in the presence of incomplete data.

Also like any statistical method, Rubin's multiple imputation approach is not a universal recipe for every incomplete-data problem. For example, for the 'in-house' type of applications, if the joint modelling of the missing-data mechanism and the substantive analysis is not difficult, both in terms of model building and implementation, then one should at least consider such a joint-modelling approach, which can be more efficient than the multiple imputation approach in terms of both statistical efficiency (e.g. avoiding the reliance on a finite number of imputations) and computational efficiency (e.g. avoiding simulation via the use of the EM algorithm).

In summary, Rubin's multiple imputation method is a powerful tool for a variety of real-life incomplete-data problems, as the three reviewed applications demonstrate. In the context of dealing with nonresponse in public-use data files (e.g. Section 4), it is a method without serious competition in terms of both generality and validity because of the unavoidable separation of the creators and the users of the database and because of the information and resource constraints on the average users for sensibly handling the nonresponse.[3,8,9] When used for other purposes (e.g. Sections 2 and 3), it is an effective addition to a statistician's toolkit because of the conceptual and implementation simplicity offered by the separation of the tasks of handling incomplete data and analysing complete data.

Acknowledgements

We thank the editor, Brian Everitt, for inviting this article for *Statistical Methods in Medical Research*, and A Gelman and JL Schafer for helpful comments. Barnard's research was supported in part by NSF Grant DMS-9705158, and Meng's research was supported in part by NSF Grant DMS-9626691. This manuscript was prepared using computer facilities supported in part by several NSF grants awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

References

1 Little RJA, Rubin DB. *Statistical analysis with missing data*. New York: John Wiley, 1987.
2 Rubin DB. Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse (c/r: P29-34). In: *ASA Proceedings of Survey Research Methods Section*, 1978: 20–28.
3 Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley, 1987.
4 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.
5 Barnard J, Rubin DB. Small sample degrees of freedom with multiple imputation. *Biometrika* 1999, in press.
6 Meng XL, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* 1992; **79**: 103–11.
7 Rubin DB, Schenker N. Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine* 1991; **10**: 585–98.
8 Rubin DB. Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* 1996; **91**: 473–89.
9 Meng XL. Multiple imputation with uncongenial sources of input (with discussion). *Statistical Science* 1994; **9**: 538–73.
10 Tu XM, Meng XL, Pagano M. The AIDS epidemic: estimating the survival distribution after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association* 1993; **88**: 26–36.
11 Lemp GF, Payne SF, Neal D *et al.* Survival trends for patients with AIDS. *Journal of the American Medical Association* 1990; **263**: 402–406.
12 Tu XM, Meng XL, Pagano M. Survival differences and trends in patients with AIDS in the United States. *Journal of Acquired Immune Deficiency Syndromes* 1993; **6**: 1150–56.
13 Volberding P, Lagakos S, Koch M *et al*. Zidovudine in asymptomatic human immunodeficiency virus infection: a controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter. *New England Journal of Medicine* 1990; **322**: 941–49.
14 Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* 1972; **34**: 187–220.
15 Prentice R, Gloeckler L. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* 1978; **34**: 57–67.
16 Dempster AP, Laird NM, 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.
17 Meng XL. The EM algorithm and medical studies: a historical link. *Statistical Methods in Medical Research* 1997; **6**: 3–23.
18 Gilks WR, Richardson S, Spiegelhalter DJ, editors. *Markov chain Monte Carlo in practice*. London: Chapman & Hall, 1996.
19 Barnard J, Du J, Hill JL, Rubin DB. A broader template for analyzing broken randomized experiments. *Sociological Methods and Research* 1998; **27**: 285–317.
20 Greene JP, Peterson PE, Du J. Effectiveness of school choice: the Milwaukee experiment. *Madison Review* 1996; 2.
21 Schafer JL. *Analysis of incomplete multivariate data*. New York: Chapman & Hall, 1997.
22 Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**: 444–72.

23 Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics* 1997; **25**: 305–27.

24 Urquhart J, de Klerk E. Contending paradigms for the interpretation of data on patient compliance with therapeutic drug regimens. *Statistics in Medicine* 1998; **17**: 251–68.

25 Robins JM. Corrections for non-compliance in equivalence trials. *Statistics in Medicine* 1998; **17**: 269–302.

26 Pocock SJ, Abdalla M. The hope and the hazards of using compliance data in randomized controlled trials. *Statistics in Medicine* 1998; **17**: 303–18.

27 White IR, Goetghebeur EJT. Clinical trials comparing two treatment policies: Which aspects of the treatment policies make a difference. *Statistics in Medicine* 1998; **17**: 319–40.

28 Goetghebeur EJT, Molenberghs G, Katz J. Estimating the causal effect of compliance on binary outcome in randomized controlled trials. *Statistics in Medicine* 1998; **17**: 341–56.

29 Smith DM, Diggle PJ. Compliance in an anti-hypertension trial: A latent process model for binary longitudinal data. *Statistics in Medicine* 1998; **17**: 357–70.

30 Rubin DB. More powerful randomization-based *p*-values in double-blind trials with non-compliance. *Statistics in Medicine* 1998; **17**: 371–86.

31 Rubin DB. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 1974; **66**: 688–701.

32 Rubin DB. Bayesian inference for causality: The importance of randomization. In: *ASA Proceedings of Social Statistics Section*, 1975: 233–239.

33 Rubin DB. Assignment to treatment groups on the basis of a covariate. *Journal of Educational Statistics* 1977; **2**: 1–26.

34 Holland P. Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 1986; **81**: 945–70.

35 Frangakis CE, Rubin DB. Addressing the invalidity of intention-to-treat tests in the presence of both treatment-noncompliance and outcome-nonresponse. Technical Report 97-FR-2, Department of Statistics, Harvard University, 1997.

36 Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–90.

37 Hirano K, Imbens GW, Rubin DB, Zhou A. Causal inference in encouragement designs with covariates. Technical report, Department of Economics, Harvard University, 1997.

38 Ezzati-Rice TM, Johnson W, Khare M, Little RJA, Rubin DB, Schafer JL. A simulation study to evaluate the performance of model-based multiple imputations in NCHS Health Examination Surveys. In: *Proceedings of the Bureau of the Census Annual Research Conference*, 1995: 257–266.

39 Schafer JL, Khare M, Ezzati-Rice TM. Multiple imputation of missing data in NHANES III [discussion on 502–10]. In: *Proceedings of the Bureau of the Census Annual Research Conference*, 1993: 459–87.

40 Barnard J. Cross-match procedures for multiple Imputation inference: Bayesian theory and frequentist evaluation. PhD thesis, Department of Statistics, University of Chicago, 1995.

41 Liu C, Rubin DB. Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika* 1998; **85**: 673–88.

42 Barnard J, McCulloch RE, Meng XL. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* 2000, in press.

43 Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* 1996; **6**: 733–807.

44 Rubin DB. Bayesianly justifiable and relevent frequency calculations for applied statisticians. *The Annals of Statistics* 1984; **12**: 1151–72.