Missing.... presumed at random: cost-analysis of incomplete data

Andrew Briggs[a,*], Taane Clark[b], Jane Wolstenholme[a] and Philip Clarke[a]
[a] *Health Economics Research Centre, University of Oxford, UK*
[b] *Centre for Statistics in Medicine, University of Oxford, UK*

Summary

When collecting patient-level resource use data for statistical analysis, for some patients and in some categories of resource use, the required count will not be observed. Although this problem must arise in most reported economic evaluations containing patient-level data, it is rare for authors to detail how the problem was overcome. Statistical packages may default to handling missing data through a so-called 'complete case analysis', while some recent cost-analyses have appeared to favour an 'available case' approach. Both of these methods are problematic: complete case analysis is inefficient and is likely to be biased; available case analysis, by employing different numbers of observations for each resource use item, generates severe problems for standard statistical inference. Instead we explore imputation methods for generating 'replacement' values for missing data that will permit complete case analysis using the whole data set and we illustrate these methods using two data sets that had incomplete resource use information. Copyright

Introduction

Clinical trials are increasingly including resource use information in addition to health outcome data in order to allow economic evaluation of health care interventions. A recent review of cost assessment of healthcare technologies in clinical trials has highlighted the handling of missing data as an issue for such cost analyses [1]. Even in a most carefully designed study, data on resource use for all patients in a trial are unlikely to be complete. However, it is rare to find any discussion of how missing data were handled in economic evaluations conducted alongside clinical trials. One exception to this is a recent evaluation of hospital at home that acknowledged

> . . . relatively few patients had a complete set of such data. Hence, mean costs for each item of resource use were calculated and then aggregated to estimate the total cost per patient. Statistical testing was therefore not possible at the level of total resource use per patient. (p1804) [2]

While we applaud the clarity with which the authors acknowledged the problem of missing data, we will argue that the chosen solution (known as available case analysis) is not optimal, precisely because it is not clear that it allows statistical testing of the differences in total cost per patient between alternatives under evaluation. The use of this method may also explain why another recent economic evaluation conducted alongside a

*Correspondence to: Health Economics Research Centre, University of Oxford, Institute of Health Sciences, Headington, Oxford OX3 7LF, UK. E-mail: andrew.briggs@ihs.ox.ac.uk

clinical trial failed to report any statistical analysis relating to differences in total cost per patient between the two trial arms despite the availability of patient-level information, instead relying on sensitivity analysis to explore the implications of uncertainty [3]. Indeed, we suspect that in most cases, health economists when confronted with missing data will use very simple methods (complete case, available case or unconditional mean imputation) to overcome the problem.

The problem of missing data is not new and has received much attention in the statistical literature as to the appropriate methods for handling missing data. While in principle missing economic data alongside clinical trials is no different to other forms of missing data, the distributional form of cost data (it is commonly highly skewed) may provide challenges for the analyst. Furthermore, since economic evaluation is commonly 'piggy-backed' onto clinical trials, there is a danger that economic variables will be considered less important by researchers responsible for data collection which could result in higher rates of missingness.

The purpose of this paper is to explore the available methods for handling missing data, with a view to highlighting the problems associated with the simplistic methods and to introduce more appropriate approaches that maintain the statistical integrity of the analysis. The next section begins with an overview of the problem of missing data and of the methods available for handling the problem. The following section then employs two examples of missing data problems. The first involves a data set on hospitalisation where just one of the variables, length of stay in hospital, suffers from the missing data problem. The second example relates to a cost analysis of patients randomised to either transurethral resection of the prostate or contact-laser revapourisation of the prostate where data were missing for a number of different resource use variables. A final section offers a discussion and an appendix is given listing some popular software packages and algorithms for conducting the analyses reported in the paper.

Methods for handling missing data

Where data on resource use information has been collected as part of a clinical trial, the cost data set will be counts of resource use for each patient in the costing part of the study. The problem of missingness arises when data are not collected or may not be available for some variables and/or for some patients. This poses a problem for economic analysis as standard statistical techniques have been designed to deal with rectangular data sets.

In this section, a general overview of the missing data problem is given. First the patterns of missingness that can occur are identified. Next the commonly employed notation for the missingness mechanism is outlined. Finally, we review the methods available for handling missing data problem in a health economic context.

Patterns of missingness

Missing data can arise in a number of ways. Univariate missingness occurs when a single variable in a data set is causing a problem through missing values, while the rest of the variables contain complete information. Unit non-response describes the situation where for some people (observations) no data are recorded for any of the variables. More common, however, is a situation of general or multivariate missingness where some, but not all of the variables will be missing for some of the subjects. Another common type of missingness is known as monotone missing data, which arises in panel or longitudinal studies, and is characterised by information being available up to a certain time point/wave but not beyond that point.

Missing data mechanisms

Little and Rubin [4] outline three missing data mechanisms:

1. Missing completely at random (MCAR). If data are missing under this mechanism then it is as if random cells from the rectangular data set are not available such that the missing values bear no relation to the value of any of the variables.
2. Missing at random (MAR). Under this mechanism, missing values in the data set may depend on the value of other *observed* variables in the data set, but that conditional on those values the data are missing at random. The key is that the missing values do not depend on the values of *unobserved* variables.

3. Not missing at random (NMAR) describes the case where missing values do depend on unobserved values.

The difference between these mechanisms is quite subtle, particularly for the first two cases of MCAR and MAR. For example, consider a questionnaire distributed to patients, in order to ascertain their use of health care resources following a particular treatment intervention, where not all the questionnaires are returned. The non-response is MCAR if the reason for failure to complete the questionnaire was unrelated to any variables under consideration. Of course, such a situation is unlikely. For example, retired patients may find more time to complete and return a questionnaire than patients who have returned to work. Also, being older on average, the retired patients may make more use of health care resources. If having conditioned on the age and retirement status of the patients in the study non-response is at random then the missing data problem is MAR. However, consider that one of the reasons for non-response is that patients are not at home, but have been taken into hospital with complications related to their original procedure. Now the missing data are NMAR since the value of the data that we do not observe is driving the reason for non-response. The methods for dealing with missingness outlined in the rest of the paper are applicable to missing information that is either MCAR or MAR.

Naïve methods for handling missing data

This section outlines the simple and *ad hoc* approaches to handling missing data that are often used. The potential problems with these approaches are highlighted and the next section moves on to a description of the more sophisticated imputation procedures.

Complete-case analysis. Complete-case analysis (CCA) or listwise deletion of cases is the default method in most statistical software packages. It involves discarding cases where any variables are missing. The advantages of using this method are that it is easy to do and that the same set of data (albeit a reduced set) is used for all analyses. However, it is inefficient in that it excludes data that are potentially informative for the analysis. Furthermore, CCA will be biased if the complete cases systematically differ from the original sample (e.g. when the missing information is in fact MAR). In practice, CCA may be an acceptable method with small amounts of missing information, but it is difficult to give definitive rules of thumb as to how much missing data may be problematic. In the case of general missingness patterns in multivariate data sets, relatively small numbers of missing data points can result in the listwise deletion of a large number of cases for a CCA, and reduce the power of an analysis.

Available-case analysis. Available-case analysis (ACA) addresses the problem of inefficiency in CCA by estimating the mean for the complete cases for each variable. The major disadvantage is that different samples are used across the analysis, i.e. the sample base varies from one variable to another since a different set of patients contribute to the estimation of different variables. This leads to problems of comparability across variables, in particular regarding the covariance structure in the data set, and may explain why in some stochastic economic evaluations analysts have not reported a statistical analysis [2,3]. Since the primary purpose of cost analysis is to calculate total cost per patient across the resource use variables, then it is clear that available case analysis will lead to the sort of problems in undertaking statistical analysis of per patient cost differences highlighted in the introduction. It should also be clear that available case analysis poses problems for standard regression-type methods that might be used when the focus of cost-analysis is the marginal effect of important covariates on cost.

Imputation methods for missing data

Imputation is where the missing data can be replaced with statistical estimates of the missing values. The goal of any imputation technique is to produce a complete data set that can then be analysed using statistical methods for complete data. The aim, therefore, is to simultaneously overcome the problems associated with both CCA and ACA. Several methods exist for imputing missing values. These are described in more detail below.

Mean imputation (unconditional means). Mean imputation is a popular, though naïve, method for replacing missing data. The mean of the observed data for each variable is calculated and substituted into every case with a missing observation for that variable. It is clear that the appeal of unconditional mean imputation lies in its simplicity. However, it is easy to see why this method is seriously flawed. Firstly, by imputing the mean value in a number of cases the estimated variance or standard deviation for that variable will be underestimated (since the imputed values do not differ from the mean or each other). Secondly, estimates of covariances and correlations are also adversely affected due to the fact that the imputed values for each of the variables are by definition unconditional. Therefore the effect of this method will be to water down the observed correlation structure of the data. Thus any further analysis such as regression analysis is questionable.

Regression (conditional) imputation. A much more promising method is to use standard regression analysis to provide estimates of the missing data conditional on complete variables in the analysis. For example, for the simple case of univariate missingness in a single continuous variable Y, we fit a regression model to explain Y by the remaining p variables represented by the vector X using the complete cases (subscripted by i):

$$Y_i = \alpha + \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i \tag{1}$$

Predicted values for the expected values of the missing cases of Y (subscripted by j) can be obtained from

$$\hat{Y}_j = \hat{\alpha} + \sum_{k=1}^{p} \hat{\beta}_k X_{jk} \tag{2}$$

It should be emphasised that the equations above could be generalised to include models for non-continuous data such as binomial or count data. Note that mean imputation from the previous section is equivalent to the simplest form of regression model with only an intercept term.

Missing data are usually multivariate and it is possible to extend the procedure of regression-based imputation from the univariate case to deal with multivariate missingness. For each missing value in the data set a model can be fitted for that variable employing the complete cases of all the other variables [5]. Where the number of variables with missing values is large, the number of models to be fitted will also be large, however, efficient computational methods (such as Little & Rubin's sweep operator) can be employed [4]. Alternatively, an iterative regression approach can be adopted [6] whereby missing values in a given variable are predicted from a regression of that variable on the complete cases of all other variables in the dataset. This process is repeated for all variables with missing values using complete cases of the other variables *including previously imputed values* until a completed rectangular data set has been generated. The imputation of missing values for each variable is then re-estimated in turn using the complete set of data and the process continues until the imputed values stop changing.

Other straightforward approaches. In some other approaches, imputations are drawn from the actual values in the data set. For example, in panel data, where data are subject to attrition, the last observation for an individual may be carried forward in time in order to complete the data set, and this approach is widely used. Little and Su [7] have suggested better methods for panel data imputation based on simple row and column fits. Another imputation method is called the *hot-deck*, as used by the US Census Bureau [8]. This method completes a missing observation by selecting at random, with replacement, a value from those individuals who have matching observed values for other variables (for matching purposes, continuous variables may need to be categorised). A more general approach to the hot-deck is to define a distance function on the basis of observed variables. A missing value is imputed based on an observed value that is close in terms of distance. One such method is *predictive mean matching* [9]. A similar approach, involves predicting propensity scores for values to be missing for individuals in the data set and then imputing missing values from complete cases with comparable propensity scores, and this method has been suggested for cost data sets with missing values due to attrition [10].

In practice, there are many methods that have been proposed for imputing missing values in order to complete data sets and it is not our intention to provide a comprehensive overview. Instead, we focus on methods that employ a formal statistical model for predicting missing values. The advantage of such methods is that they retain the statistical integrity of the analysis, allowing appropriate inference that includes uncertainty in the prediction of the missing values themselves (see the section on multiple imputation below).

Maximum likelihood approaches. The maximum likelihood (ML) approach involves formulating a statistical model and basing inference on the likelihood function of the incomplete data. We will assume that the parameters of interest are the vector of means and the covariance matrix of the variables (or cell probabilities for a multinomial model in the case of strictly categorical data) [11]. In other contexts, these parameters may be measures of effect, such as (log) odds ratios. If there were no missing data, it would be straightforward to fit parameters of the model by ML methods. If there were a univariate missing data problem, then it would be possible to factorise the likelihood in order to predict the missing data conditional only on the observed data. For the problem of multivariate missingness, the likelihood does not factorise. However, use can be made of the Expectation-Maximisation (EM) algorithm [11], which is based on a very simple premise. If we knew the parameters of the model we could estimate the missing data and if we knew the missing values we could estimate the parameters. Therefore the following iterative procedure is suggested:

1. Choose starting values for missing data points
2. Estimate the parameters
3. Re-estimate the missing values, assuming that the new parameter estimates are correct
4. Repeat steps 2 & 3 until the results stop changing.

Specifically, the EM algorithm involves iterating between an Estimation-step and a Maximisation-step. The E-step involves averaging the complete-data likelihood over the predictive distribution of the missing data given our parameters in order to provide estimates of the missing data. The M-step involves maximising the likelihood given the complete data set in order to provide updated estimates of the parameters. After convergence [12], the parameters may then be used to generate predicted values for the missing data directly. Starting values for the procedure could easily be obtained from the conditional (or even unconditional) imputation methods described above for complete cases. However, it would be prudent to run the EM algorithm with alternative sets of starting values to ensure that convergence has not occurred at a local maxima.

The multiple imputation principle

It is important to recognise that when employing any imputation method we are *estimating* a missing value that is not observed. It is straightforward to see that in the case of unconditional mean imputation, the variance of the completed variable will be too low, since the imputed means do not contribute to the variance. However, the same is true with the other forms of imputation – if the expected value of the missing data point is imputed, although this is the 'best' prediction of the missing value (in the sense of mean squared error), there will be no allowance for the uncertainty associated with the imputation process. For example, if imputations are based on a regression equation, as in Equation (2) for the simple univariate missingness example, then there will be no variation between predicted values for observations with the same values for all of the other non-missing variables. Such 'deterministic' imputation approaches [6] will therefore underestimate the variance of any estimators in subsequent statistical analysis of the imputed data set. Therefore, imputed values of missing data should include a random component to reflect the fact that imputed values are estimated (using so-called 'stochastic' imputation methods [6]) rather than treating the imputed values as if they are known with certainty.

For the regression example, two components to the uncertainty in the imputation process can be distinguished. The first component is the mean squared error from the regression which represents the between observation variability not explained by the regression model. Two approaches to including this error term are either: to select a value at random from a normal distribution with variance equal to the mean squared error from the regression; or to compute the residuals from the regression and to add one of these residuals at random to each of the imputed values from the regression. Of these two approaches, the second non-parametric bootstrap approach is probably preferred since it is straightforward to do and does not rely on the parametric assumption of normally distributed errors. The second component of uncertainty comes from the fact that the coefficients of the regression model are themselves estimated rather than known. The variance of the prediction error for each covariate pattern can be obtained from the variance–covariance matrix and, assuming multivariate normality, this component of uncertainty can also be incorporated into the stochastic imputation procedure.

Clearly, once missing values are imputed with a random component, then a complete data set will no longer be unique and the results of any analysis of will be dependent on the particular imputed values. The principle of multiple imputation uses this fact directly in order to allow estimation of variance in statistics of interest in an analysis that include representation of uncertainty in the true values of the missing information.

With multiple imputation, an incomplete data set will have the missing values imputed several (M) times, where the values to fill in are drawn from the predictive distribution of the missing data, given the observed data. Each imputed data set is then separately analysed with the desired methods for complete data. The variability in the statistic of interest across the alternative data sets then gives an explicit assessment of the increase in variance due to missing data. Thus this variance of each final parameter estimate is composed of two parts: the estimated variance within each imputed data set and the variance across the data sets.

Suppose that the statistic of interest in the analysis is given by θ. The steps in the multiple imputation procedure are then:

1. Generate M sets of imputed values for the missing data points, thus creating M completed data sets.
2. For each completed data set, carry out the standard complete data analysis, obtaining estimate $\hat{\theta}_i$ of interest and its estimated variance $\hat{\text{var}}(\hat{\theta}_i)$ for $i = 1 \ldots M$.
3. Combine the results from the different data sets. The multiple imputation estimate of θ is

   $$\hat{\theta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_i$$

   (i.e. the mean across the imputed data sets) and multiple imputation estimate of variance is

   $$\hat{\text{var}}(\hat{\theta}) = \frac{1}{M}\sum_{i=1}^{M}\hat{\text{var}}(\hat{\theta}_i) + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{i=1}^{M}(\hat{\theta}_i - \hat{\theta})^2$$

The first term on the right hand side of this equation relates to the variance within the imputed data sets, whereas the term on the far right captures the uncertainty due to the variability in the imputed values, i.e. between the imputed data sets. The term $1 + 1/M$ is a bias correction factor.

The approximate reference distribution for interval estimates and significance tests is a t distribution with degrees of freedom $v = (M-1)(1+r^{-1})^2$, [13] where r is the estimated ratio of the between-imputation component of variance (numerator) to the within-imputation component of variance (denominator).

Rubin [14] shows that the relative efficiency of an estimate based on M complete data sets to one based on an infinite number of them is approximately $(1+\gamma/M)^{-I}$, where γ is the rate of missing data. With 50% missing data, an estimate based on $M = 5$ complete data sets has a standard deviation that is only about 5% wider than one based on infinite M. Unless rates of missing data are very high, there is little advantage to using more than five complete data sets [15].

Note that while it is appropriate to average across multiple imputations for additive statistics like the mean and variance care should be taken when generating multiple imputation estimates of other quantities. For example, standard deviations and correlation coefficients should not be estimated by averaging across multiple imputations. Rather, multiple imputation estimates of such quantities should be derived from the multiple imputation variances and covariance.

Bayesian simulation methods

Markov chain Monte Carlo (MCMC) is a collection of methods for simulating random draws from non-standard distributions via Markov chains [16]. Data augmentation [17] is an iterative MCMC method for simulating the posterior distribution of the missing values in the data set given the observed values. It can be thought of as a Bayesian equivalent of the EM algorithm using simulation, with the imputation step (corresponding to the E-step) being to generate predicted values for missing data, and the posterior step (corresponding to the M-step) being to estimate the posterior distribution of the parameters given the complete data. Since data augmentation is based on simulation, it does not converge to a point estimate of the parameter of interest, rather the sequence of simulated values converges to the posterior distribution of the parameter.

Two methods, Schafer [11] and Van Buuren [18], use refinements of this methodology as discussed below and have freely available software (see the appendix).

Schafer algorithms. Schafer [11] has developed algorithms that use Bayesian iterative simulation methods to impute multiple rectangular data sets with arbitrary patterns of missing values assuming MAR. By assuming that the multivariate missing problem is distributed either as a multivariate normal for continuous variables, multinomial log-linear for categorical variables, or follows a general location model for a mixture of variables, it can be split into a series of univariate problems. MCMC methods are used to solve the multivariate case by iteration. For example, suppose that the data are multivariate normal, it is then possible to generate imputations (i.e. completed data sets) from this distribution by applying an iterative algorithm that draws samples from a sequence of univariate regressions. It is important to note that the predictors used to describe the missingness should be specified in the univariate regressions for each variable with missing data.

Van Buuren algorithm. Van Buuren [18] has applied an alternative approach that is semi-parametric in nature. As for the parametric approach above, each variable has a separate imputation model with a set of predictors that explain the missingness. In addition, an appropriate form (e.g. linear, logistic) is specified depending on the type of variable. For example, binary variables will use a logistic model. Unlike Schafer [11], this methodology does not explicitly assume a particular form for the multivariate distribution, but does assume a multivariate distribution exists and that draws from it can be generated by using MCMC (Gibbs sampling) to sample from the conditional distributions (based on the models). Although, the semi-parametric nature of this approach is very attractive, MCMC must converge to a distribution that exists and not simply alternate between isolated conditional distributions. One way of checking convergence is to observe whether the standard deviations and means of the imputed variables between iterations are free of trend [18].

Examples of missing data imputation in cost data sets

In this section, the methods of imputation for missing data outlined in the previous section are illustrated. Two data sets are employed. The first is a data set of hospital episodes in the UK prospective diabetes study (UKPDS) where data are missing on length of stay in hospital, i.e. the pattern of missingness is univariate. The second is an example of multivariate missingness in a cost analysis of either transurethral resection or contact-laser revascularisation of the prostate.

Missing length of stay in hospital

The UKPDS was a randomised controlled trial of therapies for type 2 diabetes [19]. In all, 5102 patients were recruited to the study of whom 3964 were randomised to the main comparison of conventional ($n = 1138$) or intensive ($n = 2729$) management of blood glucose. As part of the study, information on all hospitalisations was collected. A total of 7684 separate hospitalisations were recorded over the 10 year median follow-up of the trial, however, length of stay was not recorded for 1262 (16%) of these. The aim of the analysis is to be able to compare the hospital length of stay between the conventional and intensive management arms of the trial.

Data available in the hospitalisations data set are summarised in Table 1, and shows that complete data are available on the age, sex and body mass index of the patients at entry into the study, as well as specialty codes for the stay and year of study in which the hospitalisation occurred.

Note that it is immediately apparent how complete case analysis can be biased. Since the average number of days in hospital is calculated across each arm of the study, but information on length of stay is only missing for patients who had hospitalisations, the resulting estimates of 13.14 days length of stay in the conventional arm and 10.93 days in the intensive arm (see Table 3), based on complete cases only, are clearly biased downwards. Of course, it could be argued that it is appropriate to base estimates of average length of stay on complete cases from the hospitalisation data only rather than the whole UKPDS patient population. While this is true, note that such an

Table 1. Summary statistics for 7684 hospitalisations for 3069 patients of the 5102 total UKPDS patient population

Variable	Obs	(%)	Mean	SD	Median	Min	Max
Continuous							
Length of stay (days)	6422	84	8.3	14.4	5	1	540
Age	7684	100	60.6	9.3	62	26	83
Year	7684	100	7.1	4.0	7	1	19
BMI	7684	100	27.7	5.6	27	16	61
Categorical							
Sex							
Male	4398	57					
Female	3286	43					
Specialty							
Cardiology	1558	20					
Other medical	854	11					
Surgery	727	9					
Urology	633	8					
Orthopaedics	613	8					
Gastrointestinal	466	6					
Ophthalmology	437	6					
Gynaecology	338	4					
Neurosurgery	271	4					
Oncology	269	4					
All other	1518	20					
Therapy group							
Conventional	1138						
Intensive	2729						
Excluded on basis of FPG	1235						
Total patients	5102						

Note: BMI – body mass index; FPG – fasting plasma glucose; Obs – number of observations; SD – standard deviation.

approach implicitly conditions on the fact that a hospitalisation occurred and is therefore a simple version of conditional imputation.

Since length of stay was the only variable that had missing values in the data set, the (stochastic) regression based imputation method was chosen. As it is clear from Table 1 the length of stay in hospital are heavily skewed and so consideration was given to transforming length of stay before fitting the regression model. The natural log transform is often employed when data are constrained to be positive and skewed, and this transformation was indicated by the Box–Cox procedure [20]; therefore a regression model was fitted to log length of stay (employing clustering to account for within patient variability) and the results are presented in Table 2.

While a specification should be chosen to maximise the predictive power of the model, for simplicity we have used an additive specification (on the log scale) with no interaction terms. Nearly all of the coefficients were found to be significant, however, the explanatory power of the model was disappointingly low with a high residual mean squared error for the model. With concerns over the model's predictive ability, it was especially important to include a random component to the missing value predictions. This was achieved by predicting a log length of stay for the missing data values and adding both a normally distributed prediction error (estimated from the variance-covariance matrix) and a bootstrapped residual from the complete cases while still on the log scale before exponentiating to give predicted lengths of stay.

It is widely known that $E[h(.)] \neq h(E[.])$ for any non-linear transformation $h(.)$, therefore it is common to use a correction when back transforming

Table 2. Results of regression on log length of stay

Variable	Coefficient	SE	p-value
Constant	1.100	0.149	<0.001
Year	-0.021	0.004	<0.001
Age	0.010	0.002	<0.001
Male	-0.068	0.036	0.056
BMI	0.008	0.003	0.004
Specialties (relative to 'All other'*):			
Cardiology	-0.153	0.054	0.004
Other medicine	-0.237	0.055	<0.001
Surgery	-0.014	0.059	0.811
Urology	-0.409	0.079	<0.001
Orthopaedics	0.140	0.068	0.041
Gastrointestinal	-0.286	0.073	<0.001
Ophthalmology	-0.761	0.063	<0.001
Gynaecology	-0.207	0.065	0.001
Neurosurgery	0.266	0.089	0.003
Oncology	-0.205	0.109	0.061
Adjusted R^2	0.06		
MSE	0.975		

Note: SE – standard error; MSE – mean squared error; BMI – body mass index. *Only the top ten most frequently recorded medical specialties are identified in the model, the remaining specialties are grouped together to form a single 'all other' category which represents only a small proportion of the total episodes of hospitalisation.

predictions made on the transformed scale to the original scale. In particular, a nonparametric approach known as smearing is commonly employed [21]. Note, however, that by adding the bootstrapped residual before back transformation, the need for a smearing correction is obviated. This is because the expectation across imputed data points is an unbiased estimate of the expectation on the original scale due to the inclusion of a bootstrapped residual on the transformed scale.

The process of imputation described above was repeated five times to give five complete data sets for analysis by multiple imputation methods. Summary information for these data is presented in the upper part of Table 3 including the results of the CCA already introduced above. The multiple imputation analysis for these data is presented in lower section of Table 3, based on the assumption that the variable of interest is the difference in length of stay between treatment arms in the trial. Despite the clear downward bias of the CCA on the absolute estimates of length of stay in each arm, the estimated difference is not greatly affected. CCA slightly underestimates the difference in length of stay and its associated standard error relative to the multiple imputation based estimate. The ratio of the between/within data set variance is very low indicating that the uncertainty in the process of imputing the missing values is not having a major impact on the analysis.

Multivariate missingness in cost data

These data were taken from an economic evaluation performed alongside a randomised controlled trial in which 100 patients were randomised to either transurethral resection of the prostate (TURP) or contact-laser vaporisation of the prostate (Laser) [22]. All resources associated with the surgical interventions, post-operative hospital stay, community care and re-operations due to treatment failures were identified over a 24-month follow-up period, and the volumes of resources used by each patient were measured. In all, 12 categories of resource use were measured for each arm of the study; unit costs were then applied to these resource volumes to obtain costs per patient. Table 4 shows the resource costs and summary statistics for the resource volumes.

All resources were considered counts except irrigation volume (continuous), operating time (continuous), re-catheterisation (binary) and re-operation (binary). Because of skew, operating time and irrigation volume were transformed using logarithms. Re-operation is the most expensive cost, but was not prevalent with only 3 procedures performed.

Table 4 also shows the degree of missing data for each resource. Of the 53 prostate cancer patients treated by TURP, 10.4% of the data points were missing, whereas for the 47 patients treated by laser 9.6% were missing. Irrigation volume had the most missing data in each group. Adopting a case-deletion approach to the 120 (10%) missing data points results in 45 (%) of the cases being discarded. This is due to the multivariate nature of the missing values. Table 5 illustrates this and shows the relationship between missing cells and observations. Of the 45 (%) discarded above, half are missing one or two values only. More than half of the missing data points resulted from incomplete irrigation volume and community care variables. The missingness patterns of individual visit variables and within the

Table 3. Summary of multiple imputation data sets and the multiple imputation analysis for the difference in length of stay between the UKPDS treatment arms

Length of stay	*n*	Mean	SD	Min	Max
Conventional arm					
Complete cases	861	13.14	29.52	0	474
Imputed data 1	1138	18.17	33.89	0	474
Imputed data 2	1138	18.05	33.30	0	474
Imputed data 3	1138	18.10	33.66	0	474
Imputed data 4	1138	18.09	34.54	0	474
Imputed data 5	1138	17.82	33.02	0	474
Intensive arm					
Complete cases	2083	10.93	19.55	0	259
Imputed data 1	2729	15.65	30.75	0	981
Imputed data 2	2729	15.43	30.40	0	985
Imputed data 3	2729	15.87	31.63	0	971
Imputed data 4	2729	15.50	30.38	0	985
Imputed data 5	2729	15.74	31.83	0	1016
Multiple imputation analysis:					
Between arm differences		Mean	Variance	SE	*p*-value
Complete cases		2.21	1.20	1.09	0.05
Imputed data 1		2.52	1.36	1.16	0.03
Imputed data 2		2.61	1.31	1.15	0.03
Imputed data 3		2.23	1.36	1.17	0.05
Imputed data 4		2.59	1.39	1.18	0.03
Imputed data 5		2.08	1.33	1.15	0.07
Multiple imputation estimates					
Mean difference	2.41				
Mean variance	1.35				
Variance of means	0.06				
MI variance	1.41				
MI standard error	1.19				
p-value	0.04				

Note: MI – multiple imputation.

individual anaesthetic variables and operation times are correlated.

A cost analysis was undertaken to estimate the difference in cost between the Laser and TURP arms of the study employing four possible methods that could be undertaken: complete case analysis; mean imputation; MI based on sampling from parameters estimated using the EM algorithm; MI based on van Buurens MCMC (Bayesian simulation) approach. In both applications of MI, five imputations were generated.

The EM approach was implemented using S-plus routines provided by Schafer [11] and available via the web (see appendix). The assumption of a multivariate normal distribution for the EM algorithm is questionable for these data since most of the variables are counts with more than half of each being zero, and two variables are dichotomous. Indeed, our attempts to estimate parameters using the algorithm based on multivariate normality resulted in failure of the algorithm to converge, even when allowing the algorithm to run beyond the default settings implemented in Schafer's algorithm, suggesting that one or more functions of the missing values are very poorly estimated [11]. The methods for assessing and detecting convergence of the EM algorithm are discussed in detail in Section 3.3.4 of Schafer's book [11].

The continuous data were therefore dichotomised into zero and positive values and the categorical version of Schafer's EM algorithm

Table 4. Summary statistics of the 100 patients in the prostate cost data set[a]

Variable description		Summary statistics							
	Unit cost	TURP ($n = 53$)				Laser ($n = 47$)			
	£	Obs	%	Med (N)	Range (%)	Obs	%	Med (N)	Range (%)
Irrigation volume	3.51	41	77.4	11.0	1.5–55.0	30	63.8	6.0	0.0–33.0
No. of GP visits	25.47	47	88.7	0	0–5	40	85.1	1	0–12
No. of visits to practice nurse	6.90	47	88.7	0	0–3	40	85.1	0	0–10
No. of visits to district nurse	13.84	47	88.7	0	0–2	40	85.1	0	0–6
Minutes of operating time	7.87	45	84.9	38.0	7.0–90.0	44	93.6	34.5	7.0–66.0
No. of general anaesthetics	12.43	46	86.8	1	0–1	44	93.6	0	0–1
No. of spinal anaesthetics	22.08	46	86.8	0	0–2	44	93.6	2	0–2
No. of inpatient days	96.55	46	86.8	4	1–7	43	91.5	3	1–10
No. of transfusions	2.13	49	92.5	0	0–6	44	93.6	0	0–6
No. of outpatient consultations	73.55	51	96.2	1	1–2	47	100.0	1	1–4
Re-catheterisation	10.56	52	98.1	(22)	(42)	47	100.0	(14)	(30)
Re-operation	981.44	53	100.0	(1)	(2)	47	100.0	(2)	(4)
Obs with no missing cells	–	34	58.5	–	–	21	51.1	–	–
Non-missing cells	–	570	89.6	–	–	510	90.4	–	–
Missing cells	–	66	10.4	–	–	54	9.6	–	–

[a] Excludes fixed costs: TURP £93.76, Laser £398.89; Obs – number of observations; Med – Median.

Table 5. Summary of the frequency of missing data in the prostate cost data set

	Number of missing values in an observation							Missing cells
Variable	1	2	3	4	6	8	10	
Irrigation volume	17	3	2	2	2	1	2	29
GP visits			9		1	1	2	13
Practice nurse			9		1	1	2	13
District nurse			9		1	1	2	13
Operating time	2		4	1	1	1	2	11
General anaesthetic		1	4	1	1	1	2	10
Spinal anaesthetic		1	4	1	1	1	2	10
Inpatient days		3	2	1	2	1	2	11
Transfusions			2	1	2		2	7
Outpatient consulation							2	2
Catheterisation				1				1
								Total: 120
Observations missing data[a]	19	4	15	2	2	1	2	Total: 45

[a] 55 observations had no missing data.

was used to provide the parameter estimates and impute missing values. Starting values for the algorithm were the defaults provided by the algorithm, which assume equal probabilities in each of cells of the cross-classification table (see Schafer [11, Section 7.3]).

For the Bayesian simulation approach we employed, a separate imputation model was specified for each variable. The purpose of the models was to provide a set of plausible values for the missing resource data, and this involved two modelling choices: the form of the model (linear, logistic etc.) and the set of predictors that enter the model. For binary variables (e.g. re-catheterisation) we used a logistic model, for continuous variables (e.g. log operation time) we used linear regression, for the visitation and transfusion variables we used a predictive mean matching model due to the multi-modal and skew nature of the observed data, but for other count variables (e.g. number of outpatient consultations) we applied a multinomial (> 2 levels) logistic model. We proposed to confine the imputation model for each resource to include all other resource variables. This is an explicit attempt to model the MAR process, because we assuming the missingness can be explained by observed data. In other cases, auxiliary variables may be available, and it has been observed that including as many predictors in the imputation model as possible tends to make the MAR (and possibly NMAR) assumptions more plausible [18].

The Gibbs sampling algorithm was run for 150 iterations for each of 5 imputations. In general, in the presence of large amounts of missing data, convergence can be obtained in as few as 10 iterations [18]. Plots of the standard deviations and means of the imputations by iteration were free of trend, indicating that the imputation-variability had stabilised and there may be convergence.

The results of these analyses are presented in Table 6. Fixed costs have been excluded from this analysis. It is unfortunate in this particular example, that none of the missing data strategies gives a result that is significant at the conventional 5% level, however, for comparative purposes it is the *t*-ratio and *p*-value that show the strength of the evidence in favour of a cost difference between the two treatment alternatives. The CCA strategy produces a cost advantage for the Laser procedure, whilst the other strategies do not – although it would be unwise to infer too much from this given the general lack of significance. The CCA is also relatively inefficient compared to the other methods as evidenced by the large standard error for the cost difference. The unconditional mean imputation method gives the same estimate of cost difference as the ACA (not presented) but also gives an estimated standard error. Unfortunately, this may be an underestimate of the standard error of the cost difference and as such is potentially seriously misleading for statistical inference. Assuming MAR, the MI approach using Bayesian simulation could be considered to be the most reasonable approach because of its minimal assumptions. Its results are similar to the mean imputation method. The MI approach using EM produces the least absolute difference and has a standard error second only to the CCA in magnitude. Overall, the results in the non-CCA analyses are similar and, perhaps, we may have seen greater differences if there was substantially more than 10% of the data missing.

Discussion

Missing data are very common in health economic evaluations of patient-level data. The standard approach to missing information in many statistical packages is to exclude those individuals for whom data are missing from the analysis, (complete case analysis). In addition to being inefficient, this practice could lead to invalid results if the excluded group is a selective sub-sample (non-random sample) from the entire sample – a violation of the MCAR assumption. By contrast, recent cost analyses appear to have employed an available case approach to handling missing data with the consequence that standard statistical methods could not be employed despite the existence of patient-level data.

Table 6. Effect of various missing data strategies on the statistical analysis of patient-specific cost data* (£)

	Complete case analysis		Mean imputation		MI using EM parameters		MI Van Buuren MCMC	
	T	L	T	L	T	L	T	L
N	31	24	53	47	53	47	53	47
Mean	888	796	853	877	850	865	868	890
SD	222	331	230	325	253	343	265	338
Diff		91		−24		−15		−22
SE (Diff)		78		57		69		62
t-ratio*	1.16		−0.36		−0.22		−0.35	
p-value	0.26		0.68		0.81		0.72	

Note: L – Laser, T – Turp, MI – Multiple Imputation, MCMC – Markov chain Monte Carlo, EM – Expectation maximisation algorithm, *Excludes fixed costs: TURP £93.76, Laser £398.89.

This paper reviewed strategies for analysing data in the presence of missing values and illustrated some of those strategies using two examples. Our goal was not to be exhaustive in reviewing this area, but to provide an introduction to more commonly applied methods and software routines. In particular, we have focused on imputation methods that involve statistical modelling such that the uncertainty in the prediction process can be captured through the principle of multiple imputation. Thus, traditionally popular single imputation methods such as the hot deck that involve identifying a nearest neighbour to serve as a surrogate missing value have not been employed

Most of the non-CCA methodologies discussed assume that the data are Missing at Random (MAR), that is, the missing values do not depend on the values of unobserved variables. This is a more realistic assumption than MCAR. We described three broad imputation approaches: regression imputation, the use of the EM algorithm to calculate maximum likelihood estimates for use in imputing the data, and Bayesian simulation methods. Inferences associated with single imputations (i.e. single complete data sets) resulting from these methods tend to overstate precision because they omit the between-imputation component of variability. Multiple imputation, an extension of single imputation, pools more than one complete data set allowing for the uncertainty in the imputations to be appropriately reflected in the analysis.

Our examples demonstrated some fundamental ways of exploring and handling missing data. The proportion of missing data is important. It may be reasonable to perform a CCA if a small percentage of patients are missing data, but as the percentage increases there will be greater inefficiency and more chance of bias. In general, it is advisable to investigate the missing data and, in most cases, attempt to impute or fill in the missing data. In the case of univariate missingness, it may be acceptable to apply a regression imputation making adjustments for the predictions (see the first example based on missing length of stay in the UKPDS).

Most missingness is multivariate and the absence of a small percentage of data points can potentially lead to a depleted CCA. In the second example of Laser versus TURP interventions, 10% of cells were missing, but 45% of patients were removed in a CCA. It seems sensible in this case to impute 10% of the missing data in order to reinstate the extra 45% of patients. Imputation in this setting requires allowances for the correlation between variables, and strictly univariate approaches may be inappropriate. The EM algorithm was employed to estimate a covariance matrix for the variables, but only after making the variables discrete due to the problem of large numbers of zeros for individual parameters. Our preferred method was to use Bayesian simulation to create imputations, particularly the method of Van Buuren [18]. This approach requires the specification of imputation for each model and Gibbs sampling is used to develop a multivariate distribution for the data.

Bayesian methods are becoming increasingly popular and Data augmentation (DA) [5], Gibbs sampling (a particular type of DA) [16,18] and other Markov chain Monte Carlo methods are becoming more widely applied in this area. A Bayesian approach could have been used to handle missing data in the first of our examples, using for example, the WinBUGS [23] software package. In common with many other areas of statistical analysis in the absence of prior information to inform the values of missing data Bayesian and frequentist approaches are not likely to generate substantially different results. Indeed, this was precisely what we discovered when implementing a Bayesian WinBUGS regression model for the UKPDS length of stay data. This emphasises the focus of this article on the importance of handling missing data using appropriate methods rather than on the philosophical discussion of Bayesian versus Frequentist methods. Often the EM algorithm is applied to estimate information that may be used as a starting point for a Bayesian simulation procedure [5].

The presumption of this paper has been on the estimation of missing values when data are MAR. In economic evaluation alongside clinical trials it is common for repeated measures to be taken such that a monotone pattern of missingness is observed with data points on individual being observed up to a certain time point within the study but not beyond. Where this has occurred due to differential recruiting times to a study with a fixed analysis end time then observations are said to be censored. The analysis of censored cost data has been popular area of research in recent years [24–29] and the assumption of uninformative censoring underlying the method equates to the MCAR framework described in this paper.

Where data are missing due to attrition rather than due to censoring then it is unlikely that data are MCAR. Although techniques such as mixed models and generalised estimating equations can be used to make inference from data that are subject to attrition [30] without the need for imputation of missing values, the implicit assumption underlying these methods is MAR. In practice, attrition will often be linked to specific reasons for patients leaving the study early that means censoring is informative and missing values are NMAR. In such situations, these reasons underlying the attrition need to be fully explored. It may be appropriate to correct potential biases using techniques developed in econometrics sample selection literature [31] that deal with non-random drop out in panel data [32]. Alternatively, methods for informative drop out have been developed in the statistics literature for longitudinal (or panel) data [30]. In general the additional assumptions necessary for handling missing information that are NMAR can often be quite limiting and may prove to be untestable – leading to a reliance on sensitivity analysis. One approach to the problem of missing data, therefore, is to collect additional information on missing values. However, this solution only serves to emphasise that imputation methods are not a cure for a poor study design and/or a poor data collection processes. The true solution to the problem of missing data is to efficiently capture all relevant information thereby assuring that the problem does not arise.

Software

A review of software for imputing missing data may be found at www.multiple-imputation.com and in Horton [33]. Below are some of the more widely used programs:

SAS

- PROC MI (version 8.1) offers three methods for creating the imputed data sets: the regression method, the propensity score method, and the Markov chain Monte Carlo (MCMC). PROC MIANALYZE (version 8.1) is used to combine the results
- PROC MIXED (version 6 onwards) can take subjects with incomplete data into its analysis.
- PRQEX3 (Sample program – version 6 onwards) estimates multivariate missing data by sampling from a multivariate normal distribution.
- IVEWARE is a SAS-based application for creating multiple imputations.

(www.sas.com, www.multiple-imputation.com)

SOLAS 3.0 for Missing Data Analysis

- This is a commercial Windows standalone program by Statistical Solutions Limited. It provides a comprehensive set of tools to perform both single and multiple imputation.

(www.statsol.ie)

S-PLUS

- The MICE program contains S-PLUS (versions 4.5 onwards) routines for flexible generation of multivariate imputations using Gibbs Sampling. [add-on available from www.multiple-imputation.com]
- NORM, CAT, MIX and PAN is S-PLUS software (version 3.4 onwards) for multiple imputation. NORM uses a multivariate normal model. CAT uses a log-linear model for categorical data. MIX relies on the general location model for mixed categorical and continuous data. PAN is used in a panel data setting. [add-on available from www.multiple-imputation.com]
- Missing data library for S-PLUS (version 6.0). This library supports model-based missing data, by use of the EM algorithm and data augmentation algorithms. The library incorporates the Schafer algorithms and provides some exploratory missing data tools.
- Oswald is an S-plus library for the analysis of longitudinal data. It includes informative drop-out modelling and GEE. [add-on available from lib.stat.cmu.edu/S]

(www.insightful.com)

SPSS

- SPSS Missing Value Analysis® is an additional module for SPSS (version 10) that provides graphical tools to investigate missing data, and imputes missing data using the EM and regression algorithms.

(www.spss.com)

STATA

- Sg116(.1) performs hot-deck imputation for missing data.
- Sg156 performs a weighted logistic regression for data with missing values using the mean score method.
- 'Impute' procedure provides imputed values by best sub-set regression.

(www.stata.com)

Acknowledgements

A.B. holds a Public Health Career Scientist award from the UK Department of Health. T.C. is funded by an NHS Training Fellowship in Medical Statistics. An earlier version of this paper was presented at the July 1999 meeting of the UK Health Economists' Study Group held in Aberdeen. We wish to thank two anonymous referees for their constructive comments, although remaining errors are our own responsibility.

References

1. Johnston KJ, Buxton MJ, Jones DR, Fitzpatrick R. Assessing the costs of healthcare technologies in clinical trials. *Health Technol Assess* 1999; **3**(6): 1–76.
2. Coast J, Richards SH, Peters TJ, Gunnell DJ, Darlow MA, Pounsford J. Hospital at home or acute hospital care? A cost minimisation analysis. *Br Med J* 1998; **316**(7147): 1802–1806.
3. Roberts TE. Economic evaluation and randomised controlled trial of extracorporeal membrane oxygenation: UK collaborative trial. The Extracorporeal Membrane Oxygenation Economics Working Group. *Br Med J* 1998; **317**(7163): 911–915.
4. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 1987.
5. Buck SF. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J Roy Stat Soc Series B* 1960; **22**(2): 302–306.
6. Brick JM, Kalton G. Handling missing data in survey research. *Stat Meth Med Res* 1996; **5**: 215–238.
7. Little RJA, Su HL. Item non-response in panel surveys. In *Panel Surveys*, Kasprzyk D, Duncan G, Kalton G (eds). Wiley: New York, 1989.
8. Hanson RH. US Bureau of the Census (ed.). *The current population survey: design and methodology*. Technical paper No. 40. Washington: 2001.
9. Rubin DB. Statistical matching, file concentration with adjusted weights and multiple imputations. *J Business Econ Stat* 1986; **4**: 87–94.
10. Polsky D, Glick H. Estimating medical costs from incomplete follow-up data [abstract]. *Value Health* 1999; **2**(3): 229
11. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall: London, 1997.
12. Wu CFJ. On the convergence properties of the EM algorithm. *Ann Stat* 1983; **11**: 95–103.
13. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable non-response. *J Am Stat Assoc* 1986; **81**: 366–374.
14. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
15. Schafer JL. Multiple imputation: a primer. *Stat Meth Med Res* 1999; **8**: 3–16.
16. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 1990; **85**: 398–409.
17. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *J Am Stat Assoc* 1987; **82**: 528–550.
18. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999; **18**(6): 681–694.
19. UKPDS Study Group. Tight blood pressure control, risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. UK prospective diabetes study group. *BMJ* 1998; **317**(7160): 703–713.
20. Box GEP, Cox DR. An analysis of transformations. *J Roy Stat Soc Series B* 1964; **2**: 211–243.
21. Duan N. Smearing Estimate: A Nonparametric Retransformation Method. *J Am Stat Assoc* 1983; **78**: 605–610.
22. Keoghane SR, Lawrence KC, Gray AM, Chappel DB, Hancock AM, Cranston DW. The Oxford Laser Prostate Trial: economic issues surrounding contact laser prostatectomy. *BJU Int* 1996; **77**(3): 386–390.
23. WinBUGS 1.3 [*computer program*]. Spiegelhalter DJ, Thomas A, Best N *et al.* 1.3. Cambridge University: MRC Biostatistics Unit; 2000.
24. Fenn P, McGuire A, Phillips V, Backhouse M, Jones D. The analysis of censored treatment cost data in economic evaluation. *Med Care* 1995; **33**(8): 851–863.
25. Etzioni RD, Feuer EJ, Sullivan SD, Lin D, Hu C, Ramsey SD. On the use of survival analysis techniques to estimate medical care costs. *J Health Econ* 1999; **18**(3): 365–380.

26. Lin DY, Feuer EJ, Etzioni R, Wax Y. Estimating medical costs from incomplete follow-up data. *Biometrics* 1997; **53**(2): 419–434.
27. Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika* 2000; **87**(2): 329–343.
28. Carides GW, Heyse JF, Iglewicz B. A regression-based method for estimating mean treatment cost in the presence of right-censoring. *Biostatistics* 2000; **1**(3): 299–313.
29. Willan AR, Lin DY. Incremental net benefit in randomized clinical trials. *Stat Med* 2001; **20**(11): 1563–1574.
30. Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. OUP: Oxford, 1994.
31. Greene WH. *Econometric Analysis*. (2nd Edn) Macmillan: New York, 1993.
32. Leigh JP, Ward MM, Fries JF. Reducing attrition bias with an instrumental variable in a regression model: results from a panel of rheumatoid arthritis patients. *Stat Med* 1993; **12**(11): 1005–1018.
33. Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am Statistician* 2001; **55**: 244–254.