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

Jane Wolstenholme & Andy Briggs

Health Economics Research Centre,

University of Oxford

When collecting patient-level resource use data for statistical analysis, observed counts of separate categories of resources will typically be weighted by unit cost information and summed to provide an estimate of per patient total cost. However, a practical problem faced by most analysts will be that 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. In this paper we provide a non-technical overview of the problem of missing data, illustrated by reference to recent published studies and by a cost analysis of patients randomised to either transurethral resection of the prostate or contact-laser revapourisation of the prostate. 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. We argue that 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.

Acknowledgements: This paper has benefited from discussions with Dr Jouni Kuha of Nuffield College, Oxford and from a lecture course on 'Advanced Methods for Social Statistics' given by Dr David Firth, also of Nuffield College, Oxford. Of course, omissions, errors and other blunders are entirely our own responsibility.

Working Paper Presented to the Health Economists' Study Group Meeting, Aberdeen, July 1999.

1. 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 (Johnston, et al. 1999). 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.' Coast, et al. (1998:1804)

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 clinical trial failed to report any statistical analysis relating to differences in total cost per patient between the two trial arms, instead relying on sensitivity analysis to explore the implications of uncertainty (Roberts 1998). 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 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. The next section therefore begins with an overview of the problem of missing data and of the methods available for handling the problem. The third section then employs a data set with missing values relating to a cost analysis of patients randomised to either transurethral resection of the prostate or contact-laser revapourisation of the prostate in order to elucidate the methods and to demonstrate the differences in the overall analysis that can be generated by the application of different methods. A final section offers a discussion and an agenda for future research.

2. Methods for handling missing data

It is important to ascertain whether missing really is missing, In some cases missing data may be what is termed 'structurally missing', i.e. the data is missing because it is not applicable, for example, length of current employment for unemployed, or attendance at a breast screening programme for males.

2.1 The missingness mechanism

Standard statistical techniques have been designed to deal with rectangular data sets. However, inevitably some data values in the data set are not observed (i.e. missing data) and this means that standard statistical methods may not be directly applicable. Missing data can arise in a number of ways and for a number of reasons. Univariate missingness occurs when a single variable in a data set is causing a problem through missing values. 'Unit nonresponse' where for some people no data is recorded. Monotone missing data for example caused by dropout in panel or longitudinal studies resulting in variables observed up to a certain time point/wave but not beyond that point. 'Item nonresponse' or general missingness where some, but not all of the variables will be missing for some of the subjects. It is this last pattern of missingness that will be the main concern of this paper, since it is thought that this is the type of missingness most likely to be encountered by health economists with data on a potentially large number of resource use variables across many different patient subjects.

Little and Rubin (1987) outline three missing data mechanisms:

1. Missing Completely At Random (MCAR). Here missing data are random cells from the rectangular data set and bear no relation to the value of any of the variables.
2. Missing At Random (MAR). In this case, the missing data are allowed to depend on the value of the observed variables in the data set. 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 variables.

For example if one considers two variables age and number of GP visits in the last year, where the age is known, but the number of GP visits is not known for all patients. Non-

response in the number of visits is MCAR if it is not dependent on either variable, MAR if it is dependent on age but not on the number of visits and NMAR if it depends on both age and the number of visits. The methods for dealing with missingness outlined in the rest of the paper are applicable to data with MCAR or MAR non-response. It is impossible to correct for data that are NMAR without collecting further information on the non-respondents.

2.2 Methods for handling missing data

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. the non-response is in fact MAR). In practice, CCA is likely to be an acceptable method with small amounts of missing data (say where more than 90% of cases are complete). However, for 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, possibly resulting in the elimination of most of the cases.

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, then summing the means. The major disadvantage is that different samples are used across the analysis, i.e. its sample base varies from one variable to another. This leads to problems of comparability across variables, in particular regarding the covariance structure between variables in the data set. Since the 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. Due to the problems of statistical interpretation of ACA in a multivariate context ACA has been argued to be worse than CCA.

Imputation

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 dataset that can then be analysed using complete data statistical methods. 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. Indeed, unconditional mean imputation is worse than either CCA or ACA methods. There are two problems with this naïve method. 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.

Hot-decking

This can be defined as a method where an imputed value is selected from an estimated distribution for each missing value. The distribution consists of values taken from similar responding units i.e. missing values are replaced with values from respondents that are similar with respect to variables observed for both. Hot-deck imputation is very common in practice and can involve elaborate schemes for selecting units for imputation.

Sequential hot-deck – This method of imputation treats the responding and non-responding units in a sequence, and the missing value is replaced by the nearest responding value preceding it in the sequence. The advantage of sequential hot-decking is its computational simplicity. The main problem with hot-decking is that the model of imputation used is both arbitrary and implicit, resulting in poorly defined statistical properties of the method.

Last value carried forward (LVCF)

This method only applies to data where repeated measures have been made on the variable in question. The last observed value is used to fill in the missing values. The advantages of this method are that it is easy to implement and understand and also allows for complete data methods to be employed. The disadvantage is that it will produce poor results where variables are expected to change over time.

Regression imputation (conditional means)

A much more promising method proposed by (Buck 1960) estimates the mean and covariance matrix from the sample mean and covariance matrix based on the complete cases, then uses these estimates to calculate the regressions of the missing variables on the present variables, case by case (i.e provides a method based on conditional imputation). Substituting the observed values of the present variables for a case in the regression yields predictions for the missing values for that case. The following outlines the steps in a regression imputation:

1. Fit a regression model for a given variable Y explained by the remaining variables X using complete cases j, e.g. for a continuous Y can use a linear model: $Y_j = \beta_0 + \beta_1' X_j + \varepsilon_j$
2. Compute the residuals $e_j = Y_j - (\hat{\beta}_0 - \hat{\beta}_1' X_j)$ for the complete cases
3. Compute the fitted value $\overline{Y_i} = \hat{\beta}_0 + \hat{\beta}_1' X_i$ for each non-respondent i

4. Impute $\overline{Y}_i = \hat{Y}_i + e_i^*$, where e_i is a residual selected at random from the set of e_js. Alternatively, e_i^* may be generated as a normally distributed random variable with a mean of 0 and a variance of s_e^2, where s_e^2 is an estimate of var(ε) from the regression.

Where Y is not continuous, the model needs to be changed. For example, if Y is binary, a logistic regression model can be used to fit Y_j for a given X_j and the fitted values $\hat{p}_i = \hat{P}(Y_i = 1/X_i)$ computed. Imputations are generated as binomial random variates with probabilities $\hat{p}_i$.

For each missing data point, a model can be fitted for the missing variable given the observed variables for that case, using the appropriate set of complete cases. When the number of variables is large the number of different patterns of missingness and thus the number of models to be fitted may be large.

A clear advantage of a conditional mean regression based imputation is that it allows for complete-data analysis while preserving the covariance structure in the original data. While regression predictions can underestimate the variance of the variable in question (since the same combination of explanatory variables will result in the same conditional mean), this problem can be overcome by adding a random component to the imputed value using the variance of the residuals from the regression. Although the regression method can be employed for general missingness in a multivariate data structure, in practice, this is likely to require an overwhelming amount of computation of different linear regressions for each pattern of missing data.

Maximum likelihood estimation using the EM algorithm

An attractive method for handling missing data is to specify a formal statistical model for the problem of missing data and to compute maximum likelihood estimates of parameters for the model on the basis of an appropriate likelihood function given only the observed data. Although rather technical, full details of this method have been clearly laid out in standard statistical texts and articles, (Little and Rubin 1987; Schafer 1997). The basic advantage of this approach is that it can simultaneously incorporate all of the observed information in the data set, which is particularly important where patterns of missingness are general rather than monotone. Although the likelihood cannot be maximised directly,

use can be made of the **E**(stimation &) **M**(aximaisation) algorithm which formalises an old ad hoc idea (Little & Rubin, 1987):

1. Replace missing values with estimated values
2. Estimate the parameters
3. Re-estimate the missing values, assuming that the new parameter estimates are correct
4. Re-estimate the parameters, and so on, until the results stop changing.

Or in terms of the two steps of the algorithm itself: the **E** step involves estimating parameters by maximum likelihood assuming that the missing values are known (at the first stage this involves assuming some starting values for the missing data); the **M** step then assumes these parameter estimates are correct and estimates the values of the missing data. These values are then passed to the **E** step and the algorithm then iterates until the results converge.

Multiple imputation

By treating imputed values as if they are observed, both mean and regression imputation ignore the uncertainties involved in the estimation of these values, therefore underestimating the standard errors and variances of the mean and regression coefficients. The under-representation of uncertainty with single imputation can be a problem.

In order to estimate variance accurately, total variance must include a term that accounts for the amount of uncertainty involved in the imputation of missing values.

This method of multiple imputation is based on techniques developed by Rubin et al (1996, 1987 & 1991), and provides a means of estimating the variance component. Multiple imputations for the set of missing values are multiple sets of probable values; these can reflect uncertainty across a single model for nonresponse and across several models. Each set of imputations is used to create a complete data set, which can then be analysed using complete data statistics.

With multiple imputation, an incomplete data set will be filled in multiple times ($M \geq 2$), (Rubin suggests that $M = 3$ works very well (Rubin and Schenker 1991)) where the values to fill in are drawn from the predictive distribution of the missing data, given the observed data. The predictive distribution reflects the uncertainty about the missing data. Each completed data set is then separately analyzed with the desired methods for complete data. Finally the different M results will be pooled into one result. The main difficulty with multiple imputation is to draw the M imputations from the predictive distribution of the missing data given the observed data.

Because multiple imputation of missing values involves random components, each of the M completed datasets and the corresponding M sets of parameter estimates differ somewhat. This variability in the parameter estimates across multiple imputations gives an explicit assessment of the increase in variance due to missing data. This variance of each final parameter estimate is composed of two parts: the estimated variance within the iteration and the variance across iterations. The variance across iterations accounts for the fact that the missing values are not observed but estimated with uncertainty.

Steps in multiple imputation

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 $\hat{\text{var}}(\hat{\theta}_i)$.
3. Combine the results: the multiple imputation estimate of θ is $\hat{\theta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_i$ and its estimated variance is $\hat{\text{var}}(\hat{\theta}) = \frac{1}{M}\sum_{i=1}^{M}\hat{\text{var}}(\hat{\theta}_i) + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{i=1}^{M}(\hat{\theta}_i - \hat{\theta})^2$ (1)

   = mean of estimated variances $\hat{\text{var}}(\hat{\theta}_i)$ + variance of the estimates $\hat{\theta}_i$.

The first term on the left relates to the variance within the imputed data sets, whereas the term on the right hand side 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.)

3. An example of cost-analysis in the presence of missing data

3.1 The 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). 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, 13 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. Of the 53 prostate cancer patients treated by TURP, 9.6% of the data points were missing, whereas for the 47 patients treated by laser 8.9% were missing.

The aim of this example is to demonstrate the differences that would result through the application of various missing data strategies for the statistical analysis of patient-specific resource utilisation and cost data. Tables 1–3 describe the pattern of missing data: where the missing values are located, how extensive they are and whether pairs of variables tend to have values missing in different cases. The variable information and number and percentage of missing data are set out in Table 1 below. Tables 2 & 3 present pairwise missingness results for Laser and TURP respectively.

Table 1. Number and (%) missing data for TURP and laser procedures

Variable	Variable name	Number and (%) missing data	
		TURP	**Laser**
		n=53	n=47
Irrigation volume	irrig	12 (23)	17 (36)
Number of GP visits	gp	6 (11)	7 (15)
Number of visits to the practice nurse	pracnr	6 (11)	7 (15)
Number of visits to the district nurse	distnr	6 (11)	7 (15)
Length of operating time (minutes)	optim	8 (15)	3 (6)
Number of general anaesthetics	genan	7 (13)	3 (6)
Number of spinal anaesthetics	spina	7 (13)	3 (6)
Number of inpatient days	ipday	7 (13)	4 (9)
Number of transfusions	transf	4 (8)	3 (6)
Number of outpatient consultations	opcon	2 (4)	0
Re-catheterisation	recath	1 (2)	0
Re-operation	reop	0	0

Table 2. Pairwise Missingness – Laser

	irrig	gp	pracnr	distnr	optim	genan	spina	ipday	recath	transf	opcon	reop
irrig	17											
gp	22	7										
pracnr	22	7	7									
distnr	22	7	7	7								
optim	18	9	9	9	3							
genan	18	9	9	9	3	3						
spina	18	9	9	9	3	3	3					
ipday	17	9	9	9	6	6	6	4				
recath	17	7	7	7	3	3	3	4	0			
transf	17	9	9	9	6	6	6	4	3	3		
opcon	17	7	7	7	3	3	3	4	0	3	0	
reop	17	7	7	7	3	3	3	4	0	3	0	0

Table 3. Pairwise Missingness – TURP

	irrig	gp	pracnr	distnr	optim	genan	spina	ipday	recath	transf	opcon	reop
irrig	12											
gp	16	6										
pracnr	16	6	6									
distnr	16	6	6	6								
optim	17	12	12	12	8							
genan	16	11	11	11	9	7						
spina	16	11	11	11	9	7	7					
ipday	12	11	11	11	12	11	11	7				
recath	12	7	7	7	9	8	8	7	1			
transf	12	8	8	8	9	8	8	7	4	4		
opcon	12	6	6	6	8	7	7	7	3	4	2	
reop	12	6	6	6	8	7	7	7	1	4	2	0

3.2 Results of applying methods for handling missing data

A cost analysis was undertaken to estimate the difference in cost between the Laser and TURP arms of the study employing five of the imputation methods discussed in the previous section. The results of the analysis are presented in Table 4. The results show that the although less than 10% of all data points in the data set were missing, a CCA results in almost 50% of the cases being discarded due to the presence of missing values. Although, it is perhaps unfortunate that in this particular example, the CCA 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. For the ACA, the required statistical analysis of cost differences cannot be undertaken – a clear limitation of the method. In contrast to the relative inefficiency of the CCA, the unconditional mean imputation method generates a highly significant result, due to the underestimate of the standard error of the cost difference. Hot deck imputation gives a result that is close to the mean imputation method. Multiple imputation gives a result that falls between the extremes of the CCA and mean imputation methods. From Equation 1 (pg 9), the overall standard error of the cost difference decomposes into a within imputed datasets variance of 60.9 and a between imputed datasets variance (bias corrected) of 7.7.

Table 4. Effect of various missing data strategies on the statistical analysis of patient-specific cost data

	ACA		CCA		Mean Imputation		Hot deck imputation		Multiple imputation MLE[#]	
	Laser	**TURP**	**Laser**	**TURP**	**Laser**	**TURP**	**Laser**	**TURP**	**Laser**	**TURP**
N	n/a	n/a	24	31	47	53	47	53	47	53
Mean	1252	971	1195	981	1252	971	1239	963	1249	959
S.D.	-	-	331	222	325	230	332	239	343	253
Diff	280		214		280		276		290	
SE (Diff)	-		78		57		59		69	
t-ratio*	-		2.73		4.92		4.71		4.22	
P-value	-		0.008638		0.0000035		0.0000080		0.0000546	

n/a - Not applicable (each variable has different number of observed data points)

[#] Based on M=5 imputations of the complete dataset

*Assuming unequal variances

Discussion

The aim of this paper has been to introduce the rationale and methods for handling missing data in the context of economic evaluation of health care interventions alongside clinical trials. Although the data set employed as an example for this paper focused on costs, all the results reported here are directly applicable to more sophisticated methods of economic evaluation that directly involve health outcomes as well as resource use variables. It is perhaps worth noting that the methods for handling missing data in a multivariate situation can be handled independently of the statistical methods used to analyse the complete (imputed) data set once the methods have been applied. Hence it would be straightforward to apply the methods outlined in this paper to a cost effectiveness dataset with the aim of producing imputed values for missing data that would allow appropriate statistical analysis of cost-effectiveness results. Indeed, it is important that where resource use information is collected alongside clinical trials with the aim of estimating cost-effectiveness, the resulting analysis is based on standard statistical methods and does not simply present point estimates of cost-effectiveness.

Of the methods described in this paper, complete case analysis, available case analysis and unconditional mean imputation are clearly inferior methods that are associated with serious flaws that will not fail to induce poor inference should they be widely employed. It is therefore important that where the level of missing data is high health economists reject these naïve methods in favour of the more sophisticated imputation based methods. Hot decking is a popular method that has been widely employed in population surveys, probably due to its ease of use. Results for our dataset have indicated that hot-decking gives a result worryingly close to unconditional mean imputation. Furthermore, the lack of an explicit statistical model for hot-decking leaves us slightly wary of the technique. For the more formal model based imputation methods, likelihood based imputation will probably be more straightforward to implement than regression based imputation, if the pattern of missingness is general rather than monotone. However, it is clear that whichever of these is chosen, multiple imputation methods should be employed over a simple single imputation method. Although this will increase the work of the analyst, but is not a great burden since it has been argued that

very few analyses of imputed complete data sets are required (Rubin and Schenker 1991). Our results show that analysis of a single imputed dataset could result in a 10% underestimate of the standard error of the cost difference. This is because a single imputation analysis treats the imputed values as known, when in fact they are subject to uncertainty.

Of concern to some observers is that imputation methods are in the practice of 'making up data'. While this is in some sense true, we have tried to demonstrate the disadvantages of more naïve methods. Furthermore, by explicitly incorporating multiple imputations, uncertainty related to the imputed values can be appropriately quantified. Of course this argument can be taken further. Multiple imputation could be undertaken employing different models for the missing data in order to quantify uncertainty related to the model specification in the process of generating imputed values. Some exploration of this, and of the related Bayesian methods for imputing data based on a Bayesian equivalent of the EM algorithm are part of a future research agenda.

Software for missing data analysis

In the past there have been few statistical software packages with procedures (other than complete-case analysis) for dealing with missing data, however these are now emerging. SPSS versions 8 and 9 have an add-on option to their base system known as 'Missing value Analysis' for carrying out regression imputation and maximum likelihood estimation. **Website: http://www.spss.com/software/spss/base/mva1.htm**

Statistical Solutions provide a recently developed software package; 'SOLAS for Missing Data Analysis'. This software provides the following missing data techniques; mean imputation; hot-decking; last value carried forward and multiple imputation. **Website: http://www.statsol.ie/solas.html**

S-Plus functions for carrying out maximum likelihood estimation are discussed in Schafer's 1997 book *Analysis of Incomplete Multivariate Data* (Schafer 1997), and available at **http://www.stat.psu.edu/~jls/**

Buck, S. F. (1960). "A method of estimation of missing values in multivariate data suitable for use with an electronic computer." Journal of the Royal Statistical Society **B22**: 302-306.

Coast, J., S. H. Richards, et al. (1998). "Hospital at home or acute hospital care? A cost minimisation analysis." BMJ **316**(7147): 1802-6.

Johnston, K., M. J. Buxton, et al. (1999). Assessing the costs of healthcare technologies in clinical trials. London, NHS Health Technology Assessment.

Little, R. J. A. and D. B. Rubin (1987). Statistical Analysis with Missing Data. New York, John Wiley and Sons.

Roberts, T. E. (1998). "Economic evaluation and randomised controlled trial of extracorporeal membrane oxygenation: UK collaborative trial. The Extracorporeal Membrane Oxygenation Economics Working Group." BMJ **317**(7163): 911-5.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York, John Wiley and Sons.

Rubin, D. B. and N. Schenker (1986). "Multiple imputation for interval estimation from simple random samples with ignorable nonresponse." Journal of the American Statistical Association **81**: 366-374.

Rubin, D. B. and N. Schenker (1991). "Multiple imputation in health care databases: an overview and some applications." Statistics in Medicine **10**: 585-598.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London, Chapman & Hall.