Handling missing data in economic evaluation alongside randomised clinical trials: methods review and empirical applications using the net benefit framework

A. Manca
S. Palmer

Centre for Health Economics, University of York

Abstract

Missing data is potentially an extensive problem in cost-effectiveness analysis alongside randomised clinical trials, where systematic collection of both, resources use and quality of life information is required. There are several reasons for the presence of incomplete records and the validity of the analysis of data with missing values is dependent upon the mechanism associated with the missing data. Until recently the only methods widely used for analysing incomplete data have been relatively ad-hoc and have tended to focus on "removing" the missing values, either by ignoring subjects with incomplete information (e.g. case deletion) or by substituting plausible values (e.g. means, regression predictions) for the missing items. Unfortunately, ad-hoc methods have several limitations. More recently, alternative approaches which attempt to correct the major problems associated with ad-hoc approaches have been proposed. However the application of these methods have focussed on either cost or quality of life data alone. Aim of the present work is (i) to describe and discuss the quantitative methods currently available for handling missing data in economic evaluations alongside clinical trials; (ii) to illustrate - with the use of case studies - the impact on the results of the analysis from the application of different techniques to deal with incomplete observations; and (iii) to provide a graphical representation of the results of the analysis in terms of cost-effectiveness acceptability curves.

Keyword(s): Economic evaluation, RCTs, missing data

Paper presented to the UK Health Economists' Study Group Conference, September 12-14 2001, City University, London

1. Introduction

Virtually any study collecting individual-level data – regardless whether longitudinal or cross-sectional, randomised or not – could be affected to some extent by the phenomenon of incomplete observations. In economic evaluation alongside randomised clinical trials this is potentially an extensive problem, and it is not uncommon to have incomplete data in both, resource use and health outcomes. The analyst's decision with regard the method to deal (or not to deal) with missing data can be a factor influencing the results of the study. Several approaches have been suggested to handle this problem, and their appropriateness depends on practical and theoretical considerations. In this work we first review some of the more popular methods currently available for handling missing data in economic evaluations alongside clinical trials, discussing their relative advantages and disadvantages. Using two case studies we then illustrate the impact that different techniques may have on the study results. Finally, we attempt to bring these elements together into the net benefits framework, illustrating graphically the results of the analyses in terms of cost-effectiveness acceptability curves.

2. Taxonomy of the missing data problem

It is not surprising that the problem of missing data is prevalent in longitudinal studies such as randomised clinical trial with follow up assessments, in which a group of *m* individuals provides a sequence of *n* measurements at a common set of time points. Some authors (Bernhard *et al.* 1998) distinguish between *avoidable* and *unavoidable* sources of incomplete data. Avoidable sources of missing values are: (a) *methodological factors*: e.g. structure and length of the questionnaire/form used to collect data, frequency of the assessments, miscoding "not applicable" versus zero value; (b) *logistic/administrative factors*: e.g. mail questionnaires, telephone interviews, staff commitment, data entry, data quality control procedures; and (c) *patient-related factors:* e.g. patient felt too ill or too distressed, patient felt fine and just wanted to forget what she/he went through; patient was not motivated. It is therefore fair to say that whatever precaution is taken in advance there will always be some data missing. The phenomenon can be taxonomised with respect to the *mechanism* generating it, or with respect to the *form or pattern* of missing data.

With respect to the mechanism, using a widely accepted classification (Little and Rubin, 1987), it is suggested that observations may be:

(a) *Missing Completely at Random (MCAR)*

In which observations are missing for reasons unrelated to the data, and therefore the complete cases are a fully representative random sub-sample of all the cases in the original sample;

(b) *Missing at Random (MAR)*

In this case, those individuals with missing observation differ from those with complete data but, the value of the missing observations are fully predictable from the variables in the data set;

(c) *Missing Not at Random (MNAR)*

Where those cases with missing observations differ from those with complete data, due to some unobservable variable(s). This is also called *non-ignorable non-response*.

With respect to the pattern of missing data, we can have:

1) *Non-monotone* pattern which could result in
   a) *Single missing items,* e.g. one dimension in the EQ-5D questionnaire, one resource use at one point in time;
   b) *Intermittent missing whole questionnaire/form,* e.g. one of the follow-up visits is skipped but the previous and following assessments are available;

2) *Monotone* pattern, in which the data are missing due to censoring. This can be further classified as
   a) *Informative*: e.g. the patient may experience a severe deterioration of her/his health status leading her/him to abandon the study; a patient may die; etc.
   b) *Non-informative*: e.g. no specific reason can be defined for the drop out mechanism.

3. Impute or not to impute?

There are three general strategies for analysing incomplete data: *imputation, weighting,* and *direct analysis of the incomplete data* (Little and Rubin, 1990). Imputation replaces missing values by suitable estimates and then applies standard complete-data methods to the filled-in data. Weighting methods discard the incomplete cases and assign a new weight to each complete case to compensate for the dropped cases. Direct analysis of the incomplete data analyses all the data

using methods not requiring a rectangular data set. Typically, when facing the problem of having to analyse a data set with incomplete observations people have taken different positions in the past. There were those who deemed unacceptable to "make up" data, and those who – instead - adopted a more pragmatic view and "tried to do something" about it. With respect to this issue the following citation summarises the dilemma quite nicely:

> *"The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data is complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases"*
>
> (Dempster and Rubin, 1983)

There have been three historic phases in the development of imputation methods (Schafer and Rubin, 1998). During the seventies the main strategies to handle missing data were simple procedures, such as case deletion and single imputation. For a number of years the only methods widely used for analysing incomplete data were relatively ad-hoc and have tended to focus on "removing" the missing values, either by ignoring subjects with incomplete information (e.g. case deletion) or by substituting plausible values (e.g. means, regression predictions) for the missing items. More recently - thanks also the increased computational power of modern PCs - alternative approaches which attempt to correct the major problems associated with *ad-hoc* approaches have been proposed. The eighties have witnessed the development and application of likelihood-based estimate procedures, such as the EM (i.e. Estimation and Maximisation) algorithm, and during the last decade more technically advanced methods such as multiple imputation using Markov Chain Monte Carlo techniques have appeared on the scene.

4. An overview of the available strategies to handle missing data

This section provides a (very) synthesised and non-technical description of some of the more frequently used methods together with a discussion of their relative advantages and disadvantages. A more extensive discussion of these methods can be found elsewhere (Little and Rubin, 1987).

Complete Case Analysis (CCA)

If the researcher opts for analysing the complete case data set, no imputation is performed and all individuals with incomplete observations are excluded from the analysis. In longitudinal studies this can result in an unacceptable amount of observations being discarded, which in turns results in a loss of statistical power in the analysis. The assumption underlying this option is that those individuals with complete data are fully representative of those with incomplete data. In other words the researcher is assuming that the data are MCAR. If this assumption does not hold then the results of the analysis may be heavily biased. The advantage in choosing to conduct a CCA is that the researcher analyses a full rectangular set of data, using standard statistical methods. This option may sound appealing to those researchers who do prefer not to "create" data.

Unconditional mean imputation

One of the most commonly used approaches is the unconditional mean imputation where missing data in each variable are replaced by the sample mean of the observed data for that variable. This method is not recommendable for a number of reasons. First, replacing missing values by the mean corrupts the marginal distribution of the filled-in variable. Second, it changes the covariances and correlations with other co-covariates in the data set. Third, the measures of variability of the imputed data are underestimated. This may result in a too narrow standard deviation for the mean of the filled-in variable. As for the CCA, the advantage is that the researcher has a full rectangular set of data to analyse. An improvement in this method is the use of conditional mean imputation, where an estimate of the mean from the conditional distribution of the variable with missing values given the variable with complete observations is imputed. This still produces distorted estimates of variation and covariation, and better approach would be to use a stochastic regression imputation (Little and Rubin, 1990).

Last Value Carried Forward/Backward

This is an imputation procedure typically adopted in longitudinal studies, and it consists with replacing the missing value with the last available observed

measurement for the subject in question. In its *backward* version the missing value is replaced by the next available observation. One obvious potential for bias with this method is when there are no-zero time trends. It can provide biased results also if there are different rates of drop out or different time to drop out in a study (Myers, 2000).

Hot deck

This method refers to selecting at random a value from individuals with completed forms and substituting it for the individuals with a missing observation. Values are selected from individuals similar to those who have missing values on the basis of a set of covariates. Hot deck can use sophisticated modelling techniques for the selections of the data to use for replacing the missing observation. This method may be appropriate for univariate variable analysis. As for the others a major advantage of hot deck is that once the values have been filled in, the analyst has a rectangular data set.

One disadvantage, common to all the techniques described so far, is that they are all single imputation methods. It could be argued that whenever missing data are replaced by one set of imputed values, later analysis will not truly reflect missing data uncertainty. The main concern of the imputation procedures must not be to fill in the empty cells, but to incorporate in the analysis the uncertainty surrounding the true value of the incomplete observations, preserving important aspects of the data distribution and of the parameter estimates. The problem becomes worse as the rate of missing information and the number of variables in the data set increase.

EM algorithm

This is a likelihood-based method. The analyst needs (a) to specify a model for the join distribution of the variables with observed and unobserved data, (b) to compute the likelihood of the observed data under the model, and (c) estimate the parameters to maximise the likelihood. Provided that the model is correctly specified, this method generates consistent and efficient ML estimates under the MCAR and MAR mechanisms. The EM algorithm uses the following iterative procedure. Starting from a determined value of the missing observation the researcher estimates the parameters of the model; next, using these estimated parameters, the values of the

missing data are estimated. The new estimated values for the missing data are then used to estimate again the parameters of the model, and so on until two sequential parameter estimates converge. WE believe that the fact that this is a fully parametric method may be a limitation, given that it is uncommon for cost and outcomes to follow well-behaved parametric distributions.

Multivariate Multiple Imputation

Multivariate multiple imputation is a simulation-based approach to the analysis of incomplete data. Its basic feature is that of creating *m*$2 imputed data sets filling-in each of the incomplete observations with values generated from their respective predictive distributions. These m$2 data sets are analysed using standard methods, and the results from each of the *m*$2 analyses are then combined together to produce estimates and confidence intervals that incorporate the missing-data uncertainty (Little and Rubin, 1987). The multiply imputed parameter estimates are highly efficient even for very small *m* (Schafer, 1999). The method has been proposed in a parametric and in a non-parametric version, and here we focus on the latter. The non-parametric approach to multiple imputation is implemented using a prediction model based on propensity scores and the approximate bayesian bootstrap to generate the imputations. The propensity score is the estimated probability that the data is missing. The missing observations are then randomly drawn from the sub-set of observations with similar propensity score.

It should be stressed that both multiple imputation and EM algorithm are based on the assumption that data is MAR. To our knowledge there is no formal test to verify the assumption that data are MAR, and this assumption is often chosen as a starting point when data are missing (Rubin, Stern, Vehovar, 1995). Instead, a test of MCAR for multivariate data with missing data is available (Little, 1988). If there is concern that data are not MAR, it is possible in principle to run the multiple imputation procedure using a model that reflects hypothesised differences between individuals with complete data and individuals with incomplete observations. The results obtained from the two models under the MAR and non-MAR assumption can then be compared to obtain a measure of the sensitivity of the inference to the missing data process. In practice to model a non-MAR process is not a trivial task,

and it has been demonstrated that exploring the assumption of MAR against MNAR relies on strong assumptions which are not testable (Curran *et al.* 1998).

5. Case studies

In this section we apply (some of) the methods discussed so far, in order to explore the impact that these may have on the results of the economic analysis. Data from two recently conducted economic evaluations alongside RCT are used. The objective here is twofold. First, we illustrate the sensitivity of the study results to the technique used to deal with incomplete observations. Second, using the net benefit framework we provide a graphical representation of the results of the analysis in terms of cost-effectiveness acceptability curves (CEAC). In the present work the comparison is limited to the following imputation strategies: (i) complete case analysis (CCA); (ii) last value carried forward (LVCF); (iii) unconditional mean imputation and (iv) multivariate multiple imputation (generating 5 imputed data sets). Two different ways of obtaining a CEAC that incorporates missing-data uncertainty - after multiple imputation - have been employed in the two case studies. In the first one, the imputed data sets are bootstrapped separately from each other, and then combined into one overall vector. In the second case study each of the imputed data sets is analysed separately by standard methods, and the results from each analysis are then combined using the algorithm developed by Rubin and colleagues (1987, 1991). Using these combined results, we simulate a new set of observations for the mean differences in cost and QALYs based on the original study sample size.

5.1. A trial of cognitive therapy for patients with chronic depression

The first case study relates to a randomised, controlled trial of cognitive therapy in patients with chronic depression. Subjects were unipolar depressed psychiatric outpatients aged 21 to 65, who had satisfied DSM-IIIR criteria for major depression in an episode within the last eighteen months. 154 subjects were randomised to receive clinical management alone (TAU; N=77), or clinical management plus cognitive therapy (CBT; N=77). The economic analysis was undertaken from the perspective of the NHS and other agencies (e.g. social services). Non-health service expenditure (e.g. patient travel costs) and indirect costs (e.g. productivity losses incurred due to absence from work) were not considered in the analysis.

Information on health and social care utilisation was collected using an adapted version of the Client Service Receipt Inventory. The questionnaire was administered alongside the clinical assessments at four (for weeks 4-20) and then eight (for weeks 20-68) weekly intervals. The primary outcome measure used for the economic analysis was the number of relapse free days. Patients were considered to have had a relapse if residual symptoms had persisted between two successive ratings two months apart, reaching a score on the Hamilton Scale of at least 13 on both occasions combined with a level of distress or dysfunction where the withholding of additional active treatment until 68 weeks was no longer justified.

5.1.1. Missing Data Problem and Techniques

Although the proportion of complete assessments was relatively high in both groups (86% in total; 84% CBT & 89% TAU) during the entire follow-up period, the proportion of patients with complete follow-up assessments for all periods was much lower. The problem of missing data for the economic analysis was exacerbated by the frequency of assessments (11 in total) during the follow-up period. Consequently complete data for all assessment periods was only available for 99 of the 154 patients (65% in total) and for 50 (65%) and 49 (64%) of patients in each of the CBT and TAU groups respectively. Imputation methods were thus applied to the 14% of missing cost assessments to enable the analysis to incorporate the observed costs of all patients in the longitudinal study of costs, rather than the sub-set of patients with complete data.

5.1.2. Cost Analysis

Table 1 reports the results of the alternative imputation methods for the calculation of mean costs. Since costs were non-normally distributed (positively skewed), the confidence intervals were estimated using non-parametric bootstrapping methods performing 1000 replications of the original data (Efron and Tibshirani, 1993).

In the complete case analysis, mean costs per patient were significantly higher in the CBT group (mean cost difference = £1263, 95% CI = £1093 to £1431). Although the mean cost differences remained significantly higher in the CBT group using each of the 3 imputation approaches, the magnitude of this difference was lower in comparison to the results obtained using the CCA (LVCF: £778 [£357 to £1119]; mean imputation: £840 [£515 to £1111] and multiple imputation: £905 [£527 to

£1296]). Consequently the value of the incremental cost difference used in the estimation of the incremental cost-effectiveness ratio and the net-benefit statistic was affected by both the decision to impute the missing cost assessments and the choice of imputation technique.

5.1.3. Cost-Effectiveness Analysis

The clinical trial demonstrated that the number of days relapse free was significantly higher in the CBT group (mean difference = 62.67 days [9.02 to 115.52]). The incremental cost-effectiveness ratio (ICER) was thus calculated as the difference in mean cost divided by the difference in the number of days relapse free from baseline to 18-months. To avoid the problems associated with standard methods for calculating confidence intervals around ratio-based statistics, the net benefits framework was also used to explore the impact of the different approaches in relation to the net benefit statistic. The cost-effectiveness acceptability curves (CEAC) for each of the approaches were estimated in order to summarise and compare the uncertainty in the estimates of cost-effectiveness. The value of the ICER reported in Table 1 was clearly influenced by the approach to handing missing cost data. Using the CCA approach, the ICER was £20.16 per additional relapse free day. The value of the ICER fell to between £12.42 and £14.45 based on the different imputation approaches. The estimate of the ICER for CBT was most favourable using the LVCF approach and least favourable using the CCA. The ICER using the mean and MI methods lay between these approaches, although their value was closer to that obtained using LVCF as opposed to CCA.

To reflect the uncertainty in the estimates of mean costs and effects, Figures 1-4 present the scatter-plot of the mean differences in cost and relapse-free days between CBT and TAU plotted on the cost-effectiveness (CE) plane, estimated by repeated sampling as part of the bootstrapping exercise. To incorporate the results of each the 5 datasets derived using multiple-imputation, Figure 4 presents the combined results (5000 replicates in total) obtained from separately estimating 1000 bootstrap replicates for each individual dataset and combining these estimates into one overall vector. In each of Figures 1-4 there exists a high concentration of points in the NE quadrant of the CE-plane where the ICER is positive (indicating that that CBT is both more effective and more expensive than routine care alone). However, despite the similarity between CCA and the imputation methods in terms of their

overall positioning in relation to the four quadrants of the CE-plane, there are clear differences between the approaches in terms of the position and dispersion of these replicates within the NE quadrant. The position of the bootstrap replications in the imputation approaches indicates that there are large differences in both the mean and variance of the ICER depending on whether the analyst attempts to impute missing data. This finding has important implications in this example for the stochastic analysis of cost-effectiveness using either the ICER or the net benefit statistic. In each of the 3 imputation methods, the bootstrap replications are clustered around a central value that is positioned closer to the x-axis than those using CCA. This simply provides a visual summary of the results outlined in Table 1, demonstrating a lower mean cost difference using each imputation approach (NB: values on the y-axis, representing effect differences, are not affected since the imputation methods were only applied to the cost data in this example). While each of the 3 imputation methods results in a lower mean cost difference, the variance around this value is much greater in comparison with CCA. Furthermore, although the central value for each of the imputation methods is quite similar, there are clear differences in the dispersion of the bootstrap replications obtained in each approach. In particular, the added variance obtained by estimating more than one potential value for each missing data item results in a wider dispersion of replications using multiple imputation as opposed to mean imputation.

Table 1 presents the results using a net-benefits approach using three possible ceiling values of the decision-makers willingness to pay for an additional relapse-free day (λ=£10-£20). For a value of λ between £15-£20, CBT had a positive net-benefit in each of the imputation methods and a negative net benefit using CCA. Consequently the interpretation of the net benefit statistic was different between the imputation approaches and CCA, depending on the chosen value of λ. However, since none of these differences were significant at conventional levels, it would not be possible to rule out the null hypotheses of no difference using a standard frequentist interpretation. Figure 5 presents the cost-effectiveness acceptability curves (CEAC) for each approach, incorporating the uncertainty around the sample estimates of mean costs and outcomes and the uncertainty about the maximum or ceiling value for an additional unit of effect that the decision maker would consider acceptable (Van Hout et al, 1994). Since the CEAC based on the ICER and net-

benefit approaches are equivalent (Briggs, 2001) the framework is not critical to the interpretation of these curves. Although a strict frequentist interpretation of the cost-effectiveness acceptability curves is possible, the more natural way to interpret these curves is as the probability that the intervention is cost-effective. The curve shows the probability that the data are consistent with a true cost-effectiveness ratio falling below any particular ceiling ratio, based on the observed size and variance of the differences in both the costs and effects in the trial. The x-axis shows a range of values for the ceiling ratio, and the y-axis shows the probability that the data are consistent with a true cost-effectiveness ratio falling below these ceiling amounts. The position of the cost-effectiveness acceptability curve using CCA in relation to the three imputation curves clearly illustrates that the interpretation of the results are potentially highly sensitive to the decision to impute the missing data for a wide range of potential values of λ. For values of λ<£100, there exists a marked difference between the CEAC using CCA and those based on the imputation approaches. Table 2 details the probability that CBT is cost-effective based on these CEACs, for the same 3 values of λ used in the previous calculation of the net-benefit statistic. For values of λ between £15 and £20, the interpretation of the results using CCA would be that CBT does not appear to be a cost-effective use of resources. However, the exact opposite conclusion would arise based on the results of any of the alternative imputation methods. In addition, the CEAC of the imputation approaches indicate that the results are relatively robust to the choice of method used to impute the missing assessments. Consequently, the critical issue relating to the cost-effectiveness of CBT relates to the decision to impute or not, rather than in the choice of imputation method. Finally, given the high probability that CBT is effective, as the value of λ increases, the sensitivity of the results to the decision to impute diminishes. At values of λ>£100, the CEACs of the imputation approaches and the CCA converge.

5.2. Case study 2 - TVT Trial

The second case study is a cost-utility analysis comparing Tension-free Vaginal Tape (TVT™) and Burch colposuspension for the management of women with a diagnosis of “primary “ genuine urinary stress incontinence (GSI). As part of the study, patient-specific resource use data were collected on all the patients in the trial,

including length of hospital stay, time in theatre and management of compilations; resource use was costed using UK unit costs at 1999-2000 prices. Health outcomes were expressed in terms of quality-adjusted life-years (QALYs) between baseline and 6 months follow-up. QALYs were based on patients' responses to the EQ-5D health questionnaire. The analysis was undertaken adopting the UK National Health Service perspective. The clinical study enrolled 344 women diagnosed with "primary" GSI. Patients were then randomised to either Burch colposuspension (n=169) or TVT^{TM} (n=175). A total of 28 patients dropped out from the study before surgery (23 colposuspension, 5 TVT^{TM}). Of the 23 patients who did not undergo surgery in the colposuspension group, 20 withdrew their consent, 2 discontinued the study due to protocol violation and 1 patient withdrew for other reasons. Of the 5 patients who did not have surgery in the TVT^{TM} arm, 2 withdrew their consent, 2 violated the protocol, and 1 withdrew for other reasons. No differences were observed in the baseline characteristics of patients who decided to leave the study. The economic evaluation was carried out on a modified intention to treat basis including only the 316 patients who underwent surgery (colposuspension, n=146; TVT^{TM}, n=170). Resource use was recorded using a patient record form for the period in which the patient was hospitalised, and then at six months from discharge during the follow-up visit. Health outcomes were recorded at baseline, at six weeks from discharge, and during the six-month follow-up visit.

5.2.1. Missing Data Problem and Techniques

At the end of the six months period there were 214 patients with complete observations on both resource use and health outcomes. There were 53 patients with incomplete data in the TVT^{TM} group and 49 in the colposuspension group. Incomplete observations were mainly in the form of missing items. Eighty-eight women had missing items in the EuroQol questionnaire only, other 14 had missing items in the resource use data only. For consistency with the previous example we compared four approaches to handle the missing data problem: Complete Case Analysis (CCA), Last Value Carried Forward (LVCF), unconditional mean imputation (UMI), and multivariate multiple-imputation (MMI).

5.2.2. Cost-Effectiveness Analysis

Table 3 shows the results of the analysis with the four alternative imputation methods for the calculation of mean costs and QALYs. As in the previous example, the confidence intervals were estimated using the non-parametric bootstrap method (Efron and Tibshirani, 1993).

In the complete case analysis, the difference in mean costs per patient were lower in the TVT group (mean cost difference = -£303, 95% CI = -£404 to -£199). On the QALYs side, the difference in QALYs was positive although not statistically significant (mean QALY difference = +0.006, 95% CI = -0.013 to 0.024). As for the previous case study, the mean cost differences remained significantly lower in the TVT group regardless of the imputation approaches, although the magnitude of this difference was lower in comparison to the results obtained using the CCA. The mean QALY difference in the two groups was still positive using the different methods, but the results seemed to be more sensitive to the chosen imputation strategy. This aspect is reflected in the sensitivity of the incremental cost effectiveness ratio to the chosen methods. Using the CCA approach, the ICER was £50,500 per additional QALY. This value increased up to £63,250 when the mean imputation method was adopted, and decreased to £17,813 in the LVCF analysis. Finally when using MMI, the ICER was £11,500 per additional QALY. Figures 5 to 9 show the bootstrapped mean differences in cost and QALYs between TVT and Burch Colposuspension plotted on the cost-effectiveness (CE) plane. Fig. 9 illustrates the bootstrap replications for each of the after-imputation datasets. As in the previous case study, there are clear differences between the approaches in terms of the position and dispersion of the bootstrap replicates. Table 3 presents the results using a net-benefits framework using three possible ceiling values of the decision-makers willingness to pay for an additional QALY. For lambda values between £30,000 and £150,000 TVT was associated with a positive incremental net benefit (INB) in each of the imputation procedures. Figure 10 presents the cost-effectiveness acceptability curves (CEAC) obtained under each strategy. Similarly to the previous case study, these curves show the probability that the data are consistent with a true cost-effectiveness ratio falling below any particular ceiling ratio, based on the observed size and variance of the differences in both the costs and effects in the trial. The cost-effectiveness acceptability curves show that the interpretation of the results are potentially highly sensitive to the decision to impute

the missing data for a wide range of potential values of λ. In particular, the CEAC obtained after MMI seems to reflect the (additional) missing-data uncertainty - after multiple imputation. Table 4 reports the probability that TVT is cost-effective based on these CEACs, for the same 3 values of λ used in the previous calculation of the net-benefit statistic. In this particular case study the interpretation of the results does not seem to be affected neither by the decision to impute or not to impute, nor by the imputation strategy.

Conclusions

The presence of incomplete observations in economic evaluation studies using patient-level data is not uncommon. It is not possible to draw definite conclusions on whether one should impute or not to impute a dataset with incomplete data, as this depends on the missing data mechanism. However, if the researcher decides to impute the missing data - believing that these are not MCAR - then multiple imputation provides a better alternative to single imputation under the MAR assumption.

In the first part of this paper we presented a methods overview of the most commonly used strategies for dealing with missing data, briefly describing their relative advantages and disadvantages. In the second part of this work, we applied some of these methods to two data sets to explore the impact that different imputation strategies may have on the results of the analysis. Finally we attempted to challenge the methodological problem of incorporating the missing-data uncertainty - after multiple imputation in the cost-effectiveness acceptability curves. We would like to stress that this is work in progress and that we are currently exploring a number of possible ways to do this. In this paper two alternative methods have been explored.

References

Bernhard J, Cella DF, Coates AS, Fallowfield L, *et al.* Missing quality of life data in clinical trials: serious problems and challenges. *Statistics in Medicine* 1998; 17: 517-532.

Briggs A. A bayesian approach to stochastic cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care* 2001; 17 (1): 69-82.

Curran D, Bacchi M, Hsu Schmitz SF, *et al.* Identifying the types of missingness in quality of life data from clinical trials. *Statistics in Medicine* 1998; 17: 547–559.

Dempster AP, Rubin DB. *Overview*, in *Incomplete data in sample surveys Vol. 2: Theory and annotated Bibliography* Academic Press, New York, 1983.

Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman and Hall, 1993.

Little RJA, Rubin DB. The analysis of social science data with missing values. Fox J, Long JS (eds) *Modern methods of data analysis* , Sage Publications Inc., Newbury Park, CA, 1990.

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

Little RJA, A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* 1988; 83 (404):1198–1202.

Myers WR, Handling missing data in clinical trials: an overview. *Drug Information Journal* 2000; 34: 525–533.

Rubin DB, Schenker N, Multiple imputation in health care databases: an overview and some applications. *Statistics in Medicine* 1991; 10: 585-598.

Rubin DB, Stern HS, Vehovar V. Handling “don’t know” survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association* 1995; 90: 822–828.

Schafer JL, Rubin D. Multiple imputation for missing-data problems. Presented at the Joint Statistical Meetings, American Statistical Association. Dallas, TX, August 1998.

Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research* 1999; 8: 3-15.

Van Hout BA, Al MJ, Gordon GS, Rutten FFH. Costs, effects and C/E ratios alongside a clinical trial. *Health Economics* 1994; 3:309-319.

Figure 1: Scatterplot using complete case analysis (CCA)

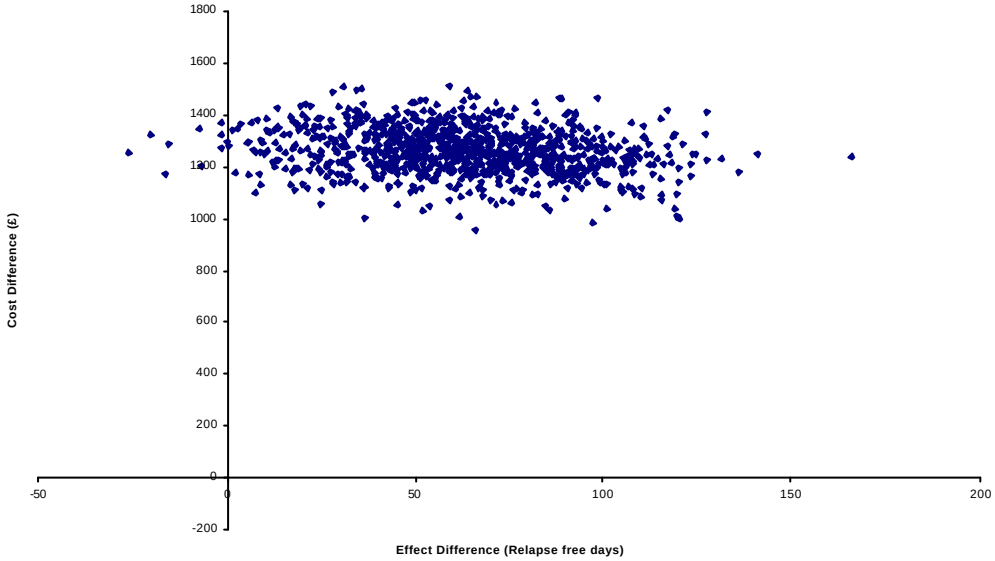

Figure 2: Scatterplot using mean imputation

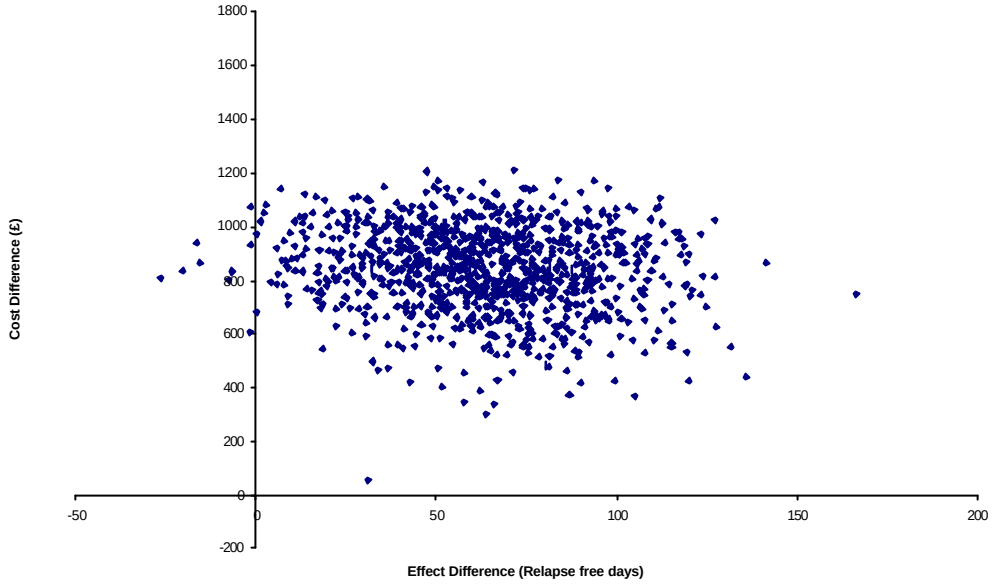

Figure 3: Scatterplot using last value carried forward (LVCF)

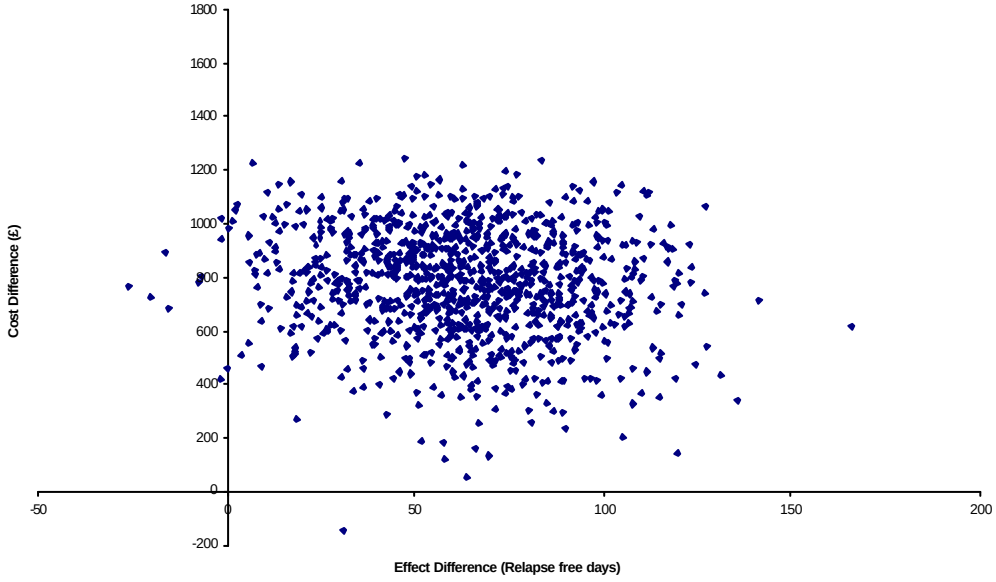

Figure 4: Scatterplot using multiple imputation (including results from 5 separate bootstraps)

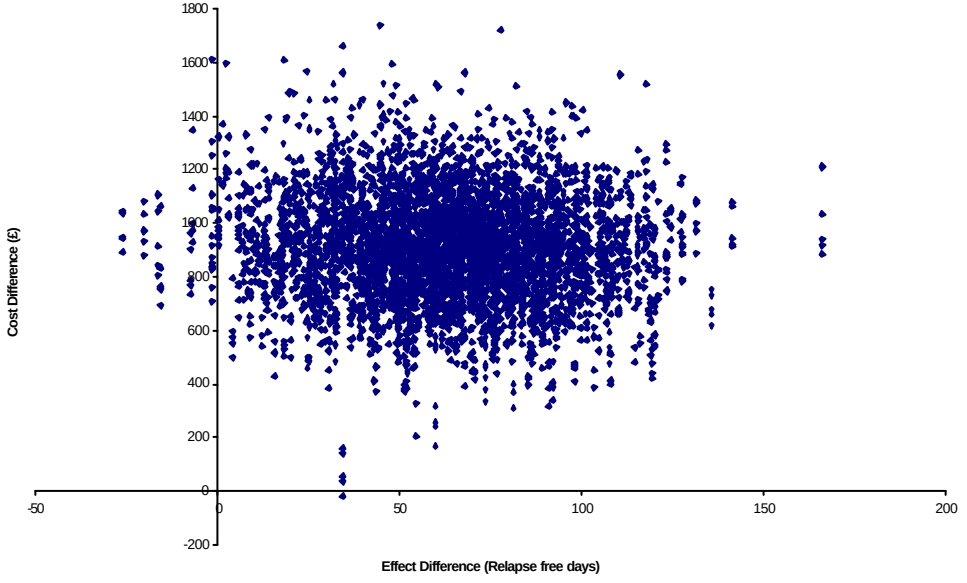

Table 1: Comparison of mean costs, ICER and net benefits using the different imputation approaches

Approach	Control Mean (s.d.)	Treatment Mean (s.d.)	Difference in means (CBT - TAU) (95% CI)	ICER	NB (95% CI)		
					λ=10	λ=15	λ=20
Complete Case Analysis	788 (375)	2051 (490)	1263 (1093 to 1431)	20.16	-637 (-1218 to -41)	-323 (-1151 to 523)	-10 (-1104 to 1089)
Last Value Carried Forward	1117 (1632)	1895 (564)	778 (357 to 1119)	12.42	-214 (-849 to 443)	100 (-773 to 992)	413 (-716 to 1526)
Mean Imputation	1069 (1231)	1909 (524)	840 (515 to 1111)	13.41	-151 (-846 to 553)	162 (-765 to 1046)	475 (-715 to 1616)
Multiple Imputation	1057 (1262)	1962 (819)	905 (527 to 1296)	14.45	-296 (-920 to 327)	17 (-827 to 892)	330 (-771 to 1449)

Table 2: Probability that treatment is cost-effective for 3 different values of lambda

Approach	λ=10	λ=15	λ=20
Complete Case Analysis	1.60%	23.30%	48.70%
Last Value Carried Forward	34.10%	62.90%	77.90%
Mean Imputation	26.50%	58.80%	76.00%
Multiple Imputation	20.58%	53.62%	73.12%

Figure 5: Cost-effectiveness acceptability curves

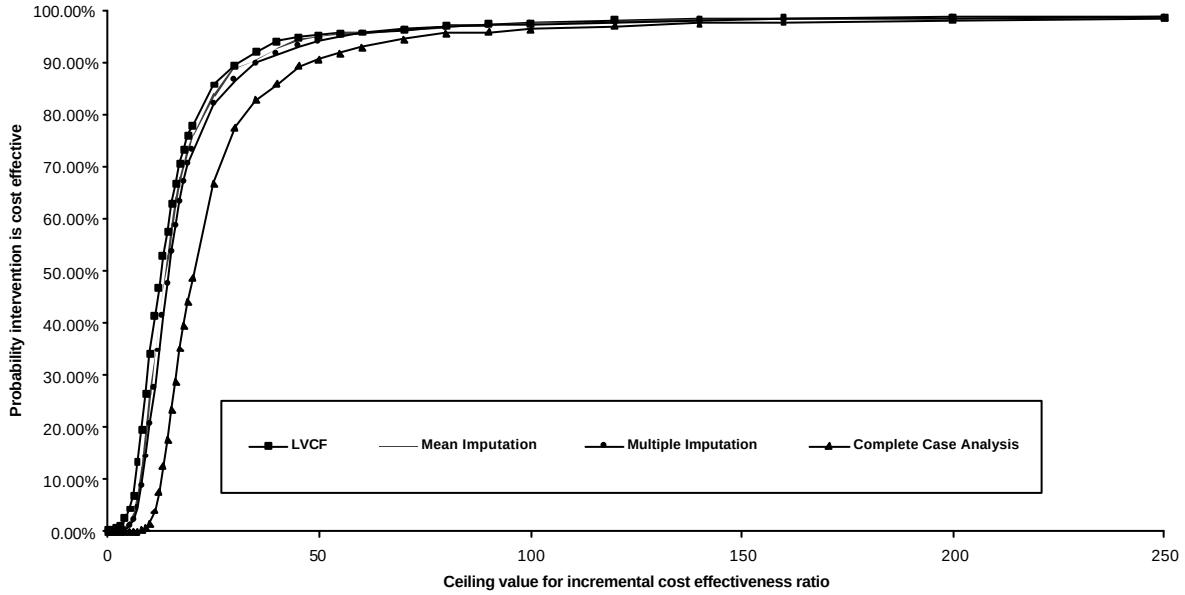

Figure 6: Scatterplot using complete case analysis (CCA)

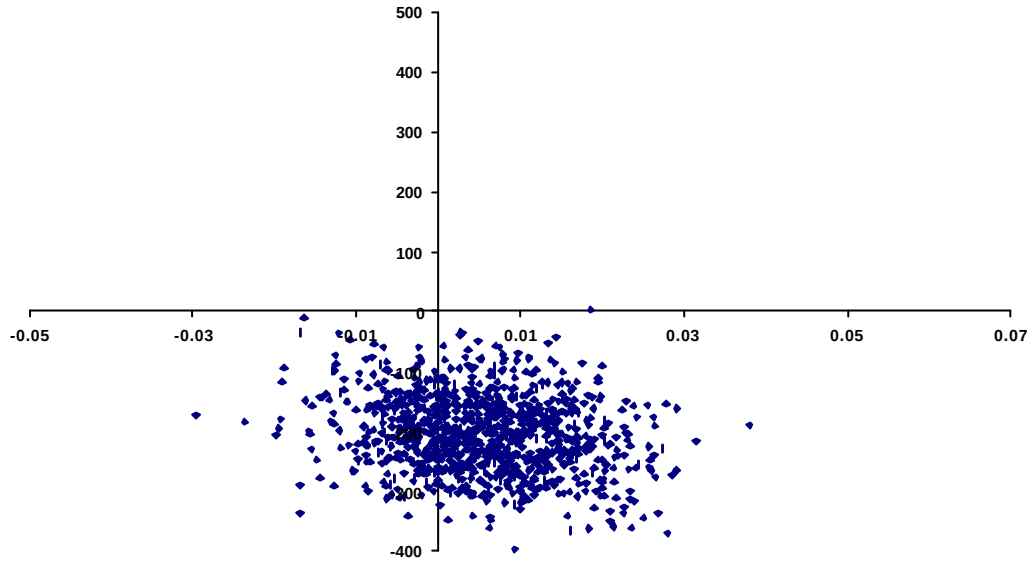

Figure 7: Scatterplot using mean imputation

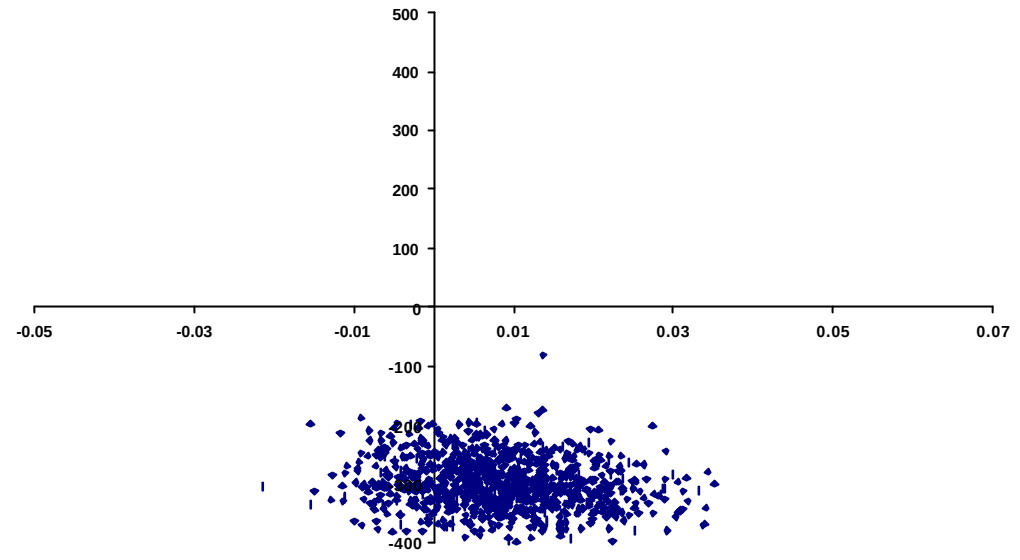

Figure 8: Scatterplot using last value carried forward (LVCF)

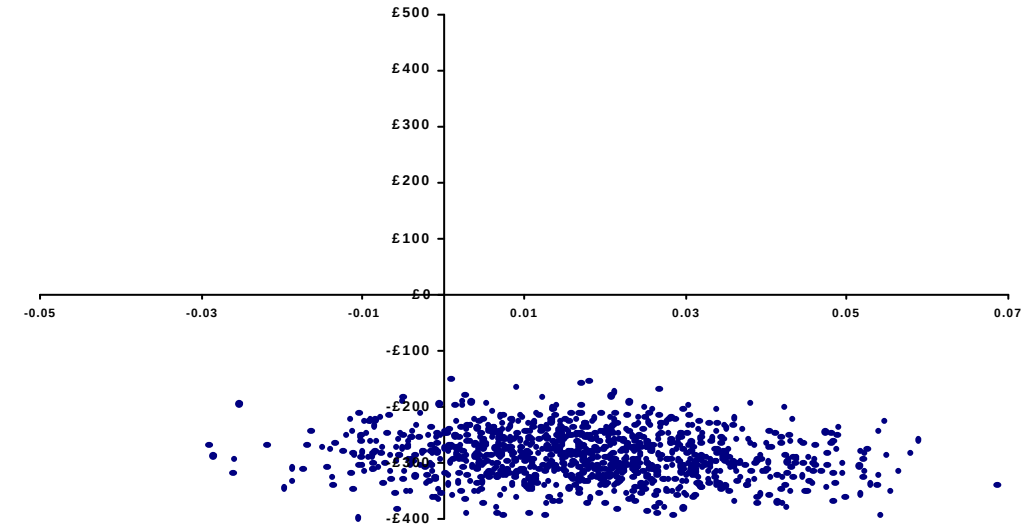

Figure 9: Scatterplot using multiple imputation (including results from 5 separate bootstraps)

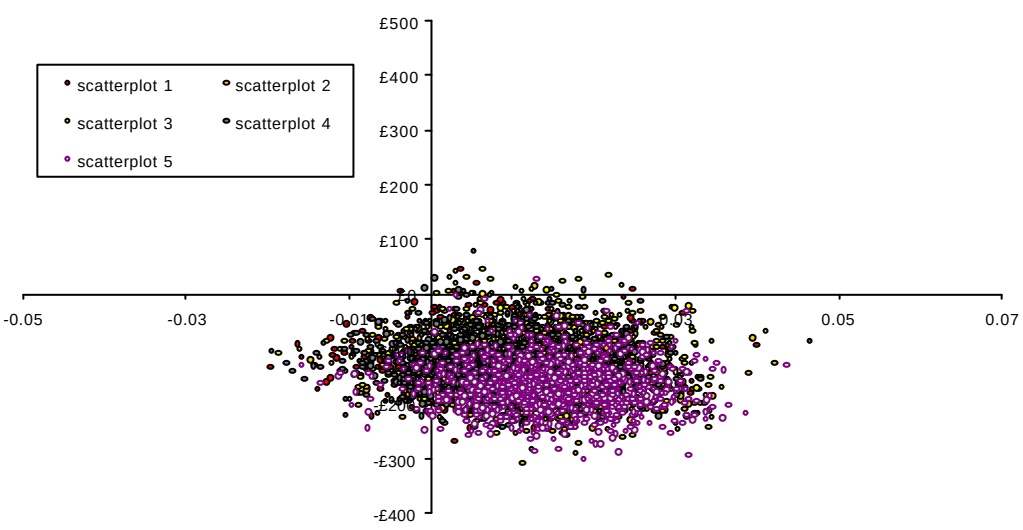

Table 3: Comparison of mean costs and QALYs, ICER and net benefits using the different imputation approaches in the TVT trial

Approach	Difference in mean costs *(TVT - Colposuspension)* (95% CI)	Difference in mean QALYs *(TVT – Colposuspension)* (95% CI)	ICER	INB (95% CI)		
				λ=30,000	λ=90,000	λ=150,000
Complete Case Analysis	- 303 (-404 to -199)	0.006 (-0.013 to 0.024)	-50500	483 (-55 to 1020)	843 (-838 to 2476)	1203 (-1607 to 3940)
Last Value Carried Forward	- 285 (-370 to -201)	0.016 (-0.011 to 0.049)	-17813	765 (-52 to 1774)	1725 (-698 to 4678)	2685 (-1341 to 7603)
Mean Imputation	-253 (-335 to -166)	0.004 (-0.008 to 0.018)	-63250	373 (110 to 1139)	613 (-221 to 2752)	853 (-570 to 4364)
Multiple Imputation	- 138 (-272 to -84)	0.012 (-0.011 to 0.020)	-11500	498 (-132 to 778)	1218 (-776 to 1972)	1938 (-1434 to 3202)

Table 4: Probability that TVT is cost-effective for 3 different lambda values

Approach	λ=30,000	λ=90,000	λ=150,000
Complete Case Analysis	95.40%	85.40%	80.10%
Last Value Carried Forward	98.40%	90.80%	90.20%
Mean Imputation	99.10%	96.20%	95.20%
Multiple Imputation	89.50%	78.00%	74.90%

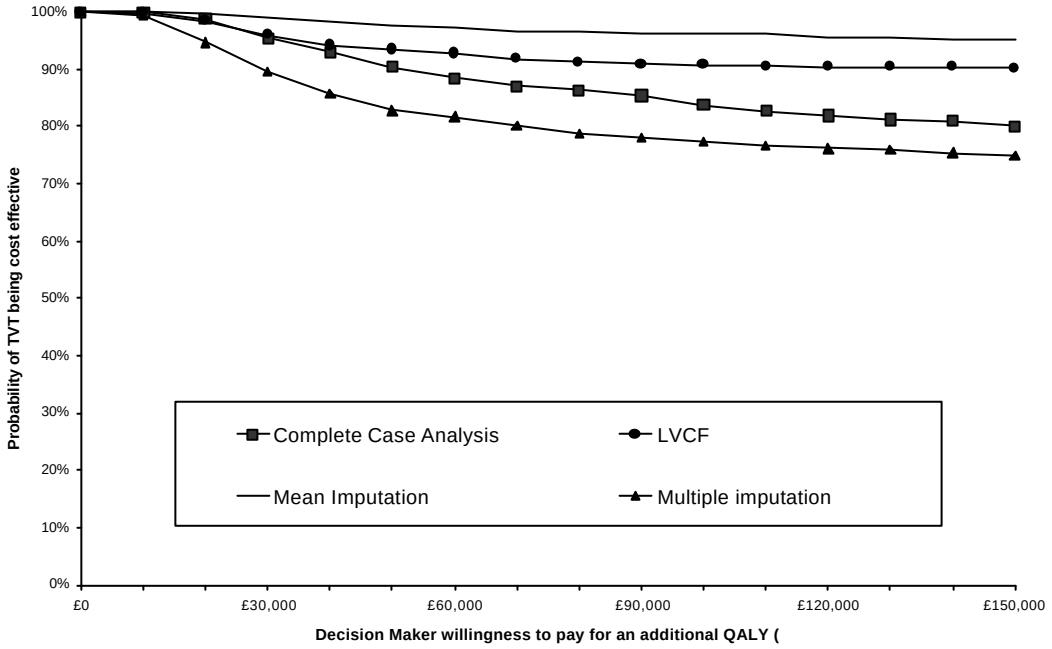

Figure 10: Cost-effectiveness acceptability curves