Combining Multiply Imputed Datasets to Produce Cost-Effectiveness Acceptability Curves Comparing Three or More Treatment Options.[1]

Susan Griffin, Centre for Health Economics, University of York (scg3@york.ac.uk).

1. Introduction

The aims of health economic evaluation are to quantify the costs and benefits of health-care interventions, and to assess their relative cost-effectiveness. The procedures by which to achieve this vary by the type of health-care intervention and the decision problem under consideration. A common issue, which needs to be addressed by any of these procedures, is that of missing data. This issue is probably most frequently addressed in trial-based economic evaluation where missing data is more apparent, given that the number of responders is less than the number of participants.

There is a range of methods available to address the problem of missing data. This paper does not seek to address the relative merits of each of these, and a useful comparison has been provided by Briggs *et al* (2003).[1] In brief, Briggs *et al*[1] conclude that in most cases, imputing the missing data is preferable to a complete- or available-case analysis. They conclude that multiple imputation procedures are superior to single imputation procedures, as they estimate the uncertainty and variance associated with imputing the missing data, rather than treating the imputed values in the same manner as the observed data. Oostenbrink *et al* (2003)[2] discuss the problem of missing data due to premature withdrawal in longitudinal clinical trials. They compare multiple imputation to several 'naive' single imputation procedures, and highlight the fact that the consequences of using inappropriate, 'naive' procedures are exacerbated when the proportion of missing data differs between the treatments under consideration. Manca *et al* (2001)[3] extend the discussion of alternative methods for analysing missing data from approaches estimating only cost data, to those estimating net benefit. Scott *et al* (2003)[4] illustrate the impact alternate methods of imputation have on the cost-effectiveness acceptability curve (CEAC).

Once the decision has been taken to use multiple imputation, it is important that all subsequent analyses properly reflect the imputation procedure. This paper

[1] Work in progress: do not quote without author's permission.

illustrates the derivation of cost-effectiveness acceptability curves or frontiers (CEACs or CEAFs) following multiple imputation in an analysis of the SoCRATES trial, a randomised controlled trial of cognitive-behavioural therapy (CBT) in early-episode schizophrenia.[5] The objectives of this paper are, firstly, to explore the implications of multiple imputation for missing data throughout a full cost-effectiveness analysis, rather than confining the discussion to only costs and the initial step of imputing the missing values. Often, papers report that multiple imputation procedures have been utilised, but fail to discuss whether the correct pooled statistics were employed, and whether they are reflected in the analysis of the results. Secondly, this paper explores the implications of multiple imputation in a cost-effectiveness analysis involving more than two comparators, with the focus on the construction of CEACs. In such an analysis, it is not possible to calculate the CEACs directly from the data, and parametric or non-parametric methods are typically employed to estimate the CEAC. This paper compares parametric and non-parametric methods for constructing CEACs, assesses the differences between them, and discusses the ability of each method to retain important characteristics of the data, such as the correct variation and correlation structure. The 'standard' non-parametric method of generating CEACs for multiway comparisons, where a large number of samples are bootstrapped from the original dataset, is discussed alongside alternative methods. This paper extends the work by Scott *et al*[4] to show that the method chosen to construct the CEAC can affect the result, and the implementation decision, over and above the chosen method of imputation.

1.1 Background

The design of the study (SoCRATES: Study of Cognitive Reality Alignment Therapy in Early Schizophrenia) was a prospective, rater-blind, randomised controlled trial. Further details about the design of the study and the clinical results have been reported in an earlier publication.[5] Patients were recruited in-hospital from three centres following an admission for first- or second-episode schizophrenia, and were randomised to CBT plus routine care, supportive counselling (SC) plus routine care or routine care alone (RC). Cost and outcome data were collected over 18-months.

The final sample for analysis consisted of 309 patients, 101 of whom were randomised to CBT, 106 to SC and 102 to RC. There were no statistically significant differences (α=0.05) between treatment groups at baseline in age, gender, ethnicity,

first versus second admission and day versus inpatient treatment; however, there were significant differences between the three centres, and so all analyses were adjusted accordingly. The sample contained patients who had a high severity of illness. This is illustrated by the fact that 83% of the sample were recruited as a first-admission, 85% were recruited as inpatients, and 38% were detained under the Mental Health Act during the acute 70-day period.

1.2 Effectiveness

For the economic analysis of the trial, the main outcome measure was selected to be a 20% reduction in PANSS (the Positive and Negative Symptom Scale)[6] total score from baseline to 18-months. Previous studies have demonstrated that a patient who exhibits a 20% reduction in PANSS total score following treatment can be clinically defined as a responder to treatment.[7-10] The choice of this measure is clinically meaningful and allows comparison with other treatments for schizophrenia. PANSS scores were recorded at baseline, 6 weeks, and 3, 9 and 18 months.

1.3 Resource utilisation and cost data

Data on length-of-stay for the initial admission and the length of time spent in CBT or SC sessions were obtained directly from patients' medical records and from the therapists in the study. Data on secondary health care utilisation were obtained from periodic case-note review of patients' medical records, and consisted of the number of outpatient attendances and visits to day hospitals. Data on utilisation of primary care and community-based services and direct non-health care costs were obtained from self-report in an interview with patients using an adapted version of the Client Service Receipt Inventory.[11] Resource use data were converted into cost data by applying the relevant unit costs obtained from national literature sources,[12] and published pricing lists.[13] Data on inpatient care and therapy utilisation were collected over the entire 18-month period. Data on utilisation of secondary health care, primary care and community-based services and direct non-health care costs were obtained at 3-, 9- and 18-months from initial admission.

1.4 The missing data problem

Missing data presented a large problem in the analysis, with only 77 patients (25%) having complete resource use data (CBT=36, SC=20, RC=21). In contrast, 308

patients (99.7%) had data on inpatient care, and full health outcome data was available for 224 patients (72%), including all those with full resource use data. Table 1 shows the pattern of missing data in the study.

Table 1: Pattern of missing data

DATA ELEMENT						N=309
Treatment	PANSS scores	Case note data	Patient questionnaire data	Inpatient use	Therapy	Number with pattern n(%)
√	√	√	√	√	√	77 (25)
√	√	√	√	√	×	8 (3)
√	√	√	×	√	√	64 (21)
√	√	×	√	√	√	30 (10)
√	√	×	×	√	√	41 (13)
√	√	√	×	√	×	1 (0.3)
√	√	×	√	√	×	2 (0.6)
√	×	√	×	√	√	33 (11)
√	×	×	×	√	√	38 (12)
√	√	×	×	√	×	1 (0.3)
√	×	√	×	√	×	7 (2)
√	×	×	×	√	×	6 (2)
√	×	×	×	×	√	1 (0.3)

√ = Observed; × = Missing

Table 1 illustrates that data from the patient questionnaire (utilisation of primary care and community-based services, direct non-health care costs and indirect costs) was the most severe singular source of missing data. It also serves to highlight how extra cases are lost when we combine data from the separate sources. If we were to ignore the resource use items on the patient questionnaire, we would gain 64 cases and have 46% complete data. Of course, any gain in the number of cases is at the expense of a more inclusive pattern of resource use. Another issue to highlight is that we have near complete data on inpatient resource use, which represents the largest component of the cost data, and the majority of which would be discarded in a complete-case analysis. Table 2 illustrates the costs and outcomes associated with complete cases with total cost, and within each of the four data sources (inpatient and therapy data are combined into one source as both are represented by a single measure over the 18-month trial period).

Table 2: Summary statistics of complete cases overall and of complete cases within the four separate sources of data.

COMPLETE CASES	**CBT** (n = 101)	**SC** (n = 106)	**RC** (n = 102)
Total Cost Data			
Number of patients (%)	**36** (36)	**20** (19)	**21** (21)
Mean total cost £ (sd)	24,024 (19,866)	17,935 (14,190)	15,381 (12,965)
Proportion responders (sd)	0.69 (0.47)	0.8 (0.41)	0.86 (0.36)
1. Inpatient and therapy data			
Number of patients (%)	**94** (93)	**88** (83)	**102** (100)
Mean cost £ (sd)	18,469 (18,755)	18,841 (17,618)	15,662 (17,883)
Proportion responders (sd)	*not complete*	*not complete*	*not complete*
2. Case note data (excluding inpatient data)			
Number of patients (%)	**65** (64)	**66** (62)	**59** (58)
Mean cost £ (sd)	1,634 (2,193)	1,594 (2,297)	1,714 (2,891)
Proportion responders (sd)	*not complete*	*not complete*	*not complete*
3. Patient questionnaire data			
Number of patients (%)	**46** (46)	**40** (38)	**31** (30)
Mean cost £ (sd)	2,861 (1875)	2,650 (1697)	2,850 (2628)
Proportion responders (sd)	0.72 (0.46)	0.88 (0.33)	0.87 (0.33)
4. Outcome Data			
Number of patients (%)	**75** (74)	**79** (75)	**70** (69)
Mean cost £ (sd)	*not complete*	*not complete*	*not complete*
Proportion responders (sd)	0.72 (0.45)	0.76 (0.43)	0.69 (0.47)

sd = standard deviation

Table 2 also serves to illustrate the fact that a complete-case analysis, including only patients with complete cost and outcome data, is unlikely to be representative of the complete patient population. The proportion of responders by treatment in the complete-case analysis was 0.69, 0.80 and 0.86 for CBT, SC and RC respectively, implying that RC was the most effective treatment strategy. However, if we examine the proportion of responders by treatment of all those with complete outcome data (72% of the full sample compared to 25% for the complete-case analysis), the inference with respect to RC is reversed, with proportions of 0.72, 0.76, and 0.69 for CBT, SC and RC respectively.

2. Method

Data were imputed using SOLAS version 3.0[14] and the propensity score method. By this method, the data are grouped into five sets using the propensity score, which describes the likelihood of the data being missing and is calculated using specified covariates. These covariates are the same for all patients, and must be present in each patient in order for a missing item to be imputed. Data are implicitly assumed to be 'missing at random'. This means that the missing data items may be related to observed variables (covariates) but not to the missing items themselves. In clinical trials, dropout frequently involves reasons closely related to the outcomes being assessed. When dropout depends only upon any or all outcomes prior to dropout, the data is 'missing at random'.[15] Table 2 indicates that missing data were typically from patients with higher costs and lower benefit from treatment, which may indicate patients with more severe disease. For example, in patients receiving SC or routine care alone, the total cost in the complete case analysis underestimates even the single, inpatient resource use component of cost (1), which was one of the most fully observed items. For those same patients, the complete-case analysis overestimates the benefits of treatment, as indicated by the analysis of the full outcome data (4). So long as the missing data relate to data recorded before dropout, the missing data may be considered missing at random. If this assumption is violated, the results of the imputation will be invalid, but as yet there is no test to determine if data truly are missing at random, and it is left to the judgement of the analyst.

Missing data are imputed by drawing from the pool of available cases that correspond to the propensity score of the missing case. Further detail is provided in Oostenbrink *et al* (2003).[2] An important point to note is that the propensity score method in SOLAS[14] requires that the missing data form a monotone pattern. This means that it must be possible to order the variables to be imputed in such a way that each variable is at least as observed (in the same participants) as the next variable to its right. The multiple imputation procedure results in five datasets[2], between which the imputed data points may vary. These datasets can be analysed to compute pooled statistics of interest for costs and effects that account for the uncertainty surrounding the imputed data points.

---

[2] When multiple imputation is used, any number of datasets can be produced. Five is the default number of datasets in SOLAS, and there is typically little efficiency gain in producing more than five. For more discussion see Schafer (1997) and Schafer (1999).

Missing data on resource utilisation were imputed with centre, treatment group and the (log) duration of untreated psychosis as covariates. The (log) duration of untreated psychosis was found to be a significant predictor of costs, and probably proxies a greater demand for resources. Baseline resource use had not been recorded. Missing data on PANSS score were imputed with centre, treatment group and baseline PANSS score as covariates. One patient in the routine care group did not have a recorded PANSS score at baseline and this patient was therefore excluded from the analysis.

2.1 Cost-effectiveness

The pooled mean of the cost and proportion of responders for each treatment strategy can be used to estimate incremental cost-effectiveness ratios (ICERs) in the standard manner.[16] In this example, the ICER represents the cost per additional responder to treatment. The decision of whether a treatment can be considered cost-effective will be determined by the maximum amount society is willing to pay for an additional responder to treatment, where this is defined as a 20% reduction in PANSS total score in patients with early-episode schizophrenia.

To represent the uncertainty in the cost-effectiveness of each treatment, CEACs or CEAFs are presented, which show the probability that each treatment is more cost-effective than all the other treatments under consideration for a range of values of willingness to pay (WTP) for an additional responder.[17, 18] These were constructed using the net benefit approach.[19] The CEAF describes the uncertainty around the optimal treatment strategy at each value of WTP, i.e. it describes the uncertainty around the treatment strategy with the highest expected net benefit, for each value of WTP. In the presence of a symmetrical distribution of net benefits, the optimal treatment strategy using the CEAC and CEAF will be identical. However, when the distribution of net benefits is skewed, the treatment with the highest probability of being cost-effective is not necessarily the treatment with the highest expected net benefit. The optimal treatment decision should be made on the basis of expected net benefits, and so under these circumstances, the CEAC cannot be used to inform the treatment decision, and the CEAF represents a more informative approach.

When constructing CEACs and CEAFs, dominated or extendedly dominated strategies need to be included, as the uncertainty around the optimal treatment strategy depends on all of the alternatives. The pooled statistics can be used directly

to calculate p-values for a pair-wise comparison of treatments, but the same is not possible for a three-way comparison. The standard, non-parametric method for constructing CEACs comparing three or more treatment options is to bootstrap a large number of samples from the original data, in order to plot the proportion of times each treatment represents the maximum average net benefit for a range of values of WTP. This is suitable when there is only a single dataset to be analysed. However, following multiple imputation, the analyst is left with M ($M > 1$) datasets which must be analysed together in an appropriate manner which reflects the imputation procedure.

2.2 Correlation

Resource utilisation and health benefits are likely to be correlated in some way. For example, those patients who experienced relatively less benefit from the treatment may have been the more difficult cases to treat, perhaps displaying higher co-morbidity, and therefore incurring higher levels of resource use. Alternatively, those patients who experienced most benefit may have made most use of healthcare resources and therefore be relatively more expensive. It is important to retain this correlation, both when imputing missing data and when constructing the CEACs.

If missing data on resource use and health benefits are imputed simultaneously, the same available case will be used to provide estimates of both missing items in each imputation. However, there are several reasons why resource use and health benefit may be imputed separately. They may be expected to have different covariates in the imputation model, especially in the case that baseline costs or effects are included. Where resource use and effects are collected separately, as was the case in the SoCRATES trial, the pattern of missing data may differ. If a monotone pattern of missing data is then required, individual resource use items and health benefits may have to be imputed separately, which allows a different available case to inform each item. Alternatively, all of the items could be imputed together if data that causes a non-monotone pattern is discarded. Following separate imputation it may not be possible to combine the five datasets of costs and effects to form five complete datasets in such a way that the original correlation structure is preserved. If the datasets are simply paired in order of imputation, then this may introduce spurious correlation from the arbitrary pairing of separately imputed cost and outcome values.

Techniques exist to address some of the issues highlighted in the paragraph above. For example, the use of seemingly unrelated regressions in the imputation procedure would allow different covariates to be used in the prediction of costs and health outcomes, whilst also allowing for correlation between the error structures in the two equations. This paper does not seek to address the issue of correlation in the choice of more sophisticated multiple imputation procedures. Instead, the aim is to provide a practical example of the possible steps to maintain the correlation structure in a dataset in which the imputation procedure used may require separation of resource use and health outcome items. The subsequent sections discuss four alternate (parametric and non-parametric) methods by which to construct CEACs from multiply imputed datasets, and the possibility within each method of including the correlation between costs and outcomes.

2.3 Non-parametric methods

As mentioned earlier, it is possible to bootstrap repeated samples from the data and to then use these to construct CEACs comparing all the options simultaneously when there is only one dataset to be analysed. However, given that we now have 5 datasets, in order to bootstrap the data we would need to either stack the datasets to form one large dataset, or bootstrap each dataset separately and then combine the results. Taking the first of these options, this will result in an underestimate of the variance about the point estimates of interest. The correct pooled variance, Var_p, from M multiply imputed datasets has two components, the between-imputation variance, and the within imputation variance. This can be written as:

$$(1)\quad Var_p = (1 + M^{-1})B + U$$

B is the between-imputation variance given by:

$$(2)\quad B = (M - 1)^{-1} \sum(\mu_i - \mu_p)^2 \qquad i = 1 \text{ to } M$$

and U is the within imputation variance given by:

$$(3)\quad U = M^{-1}\sum s_i^2 \qquad i = 1 \text{ to } M$$

p denotes the pooled estimate from the M datasets, and i denotes datasets 1 to M, μ denotes the mean and s denotes the standard deviation. When the datasets are stacked, the variance is estimated by

$$(4)\quad V = \frac{\sum\left(x_j - \mu\right)^2}{\left(N-1\right)}$$

for a sample of $j = 1$ to N, and $N = Mn$ if n is the size of the original sample. It can be shown that μ is equal to μ_p.

This formula (4) is akin to replacing μ_i with μ_p in the above formulae for B and U (where it is used to calculate s_i), thereby ignoring any variation in the means of the multiply imputed datasets, and using $Mn - 1$ degrees of freedom in the denominator where

$$(5)\quad d.o.f = \left(M-1\right)\left[1 + \frac{1}{M+1}\frac{U}{B}\right]^2$$

would be more appropriate. Bootstrapping the stacked datasets forms Method 1 for analysing the datasets.

The alternative to bootstrapping the stacked datasets is to bootstrap each dataset individually then pool the corresponding bootstrapped estimates to compute one set of pooled bootstrapped estimates. This way, the correct formulae can be used to retain the between imputation variance, and correctly estimate the within imputation variance. However, although the correct variance can be calculated, the uncertainty caused by imputing the data is still not incorporated into the CEACs, because they are constructed directly from the bootstrapped results, and do not employ the pooled estimates of the variance. Bootstrapping each dataset separately and then combining the results forms Method 2 for analysing the datasets. Under this second method, the bootstrapped results from which the CEACs are constructed are pooled means calculated from five sets of bootstrapped samples. This is in contrast to Method 1, where the bootstrapped results are samples drawn directly from the stacked dataset. Methods 1 and 2 are both nonparametric methods, and are expected to produce similar CEACs as both will not contain the between-imputation variance. The impact on the CEACs of ignoring the between-imputation variance is unclear. The CEACs must sum to one at each value of WTP, meaning that the uncertainty around every treatment strategy cannot be increased. The uncertainty around certain treatment strategies may even be reduced as a consequence of this omission, but either

way the uncertainty reflected in the CEACs is unlikely to be representative of the true uncertainty around each treatment strategy.

Within Methods 1 and 2, the correlation in the *imputed* datasets may be maintained by setting the random seed to a fixed number, so that the random picks used to generate a bootstrapped sample are the same each time. This ensures that any patient selected to provide an estimate of cost in the bootstrapped sample is also selected to provide an estimate of health benefit. However, as we mentioned earlier, the correlation in the imputed datasets could be a product of the imputation procedure itself as well as the true correlation structure. If the true correlation structure is known, it is possible for it to be re-imposed on the bootstrapped samples. This is done by re-ordering the bootstrapped estimates of costs and health benefits to reflect the desired structure.

2.4 Parametric methods

The pooled point estimates provide information for parameterising distributions of costs and health benefits. By using these estimates to define a distribution for costs and health benefits, it is possible to repeatedly simulate the sample from the distributions, in order to obtain data equivalent to the bootstrapped information, with which to construct the CEACs. The defined distribution of costs or health benefits is informed by pooled estimates of both the mean and variance, and therefore will include the between-imputation variance. This forms Method 3 for constructing CEACs, conducted with Crystal Ball[20], an Excel[21] add-in. Crystal Ball allows you to define the correlation between distributions, and the subsequent random picks from each distribution are then ordered to reflect this correlation. Again, one could enter the true correlation structure if it is known.

It is possible to embed the true correlation in the multiple imputation process by imputing net benefits[19], and this forms Method 4 for analysing the datasets. It was decided that Method 4 would be combined with Method 3 to produce CEACs, by parameterising a distribution of net benefits, rather than bootstrapping from the multiply imputed datasets. By imputing net benefits, the pool of available cases from which missing data items are drawn is reduced to only those patients with full cost and outcome data, from which net benefits are calculated. It also restricts the reporting of results to only summary statistics about the net benefit of each treatment strategy. It is necessary to impute net benefits for a suitable range of values of WTP.

Depending on the method of imputation, this may be computationally taxing, or costly in terms of time.

3. Results

Table 3 displays the summary statistics from the multiply imputed datasets. There is a large amount of variation in costs in particular, and this contributes to the fact that there are no significant differences in costs or health benefits between the three treatment strategies ($\alpha = 0.05$). The process of multiple imputation increases the variation around the point estimates of interest, but this is offset with a gain in precision from the inclusion of more cases.

Table 3: Summary pooled statistics from the 5 multiply imputed datasets

MULTIPLE IMPUTATION	**CBT** (n=101)	**SC** (n=106)	**RC** (n=101)
Mean total cost £ (sd)	23,298 (18,913)	22,992 (17,412)	20,883 (18,216)
Median total cost (I.Q.R.)	19,146 (9,758 – 28,425)	16,945 (10,852 – 28,347)	14,309 (9,577 – 24,225)
Proportion responders (sd)	0.71 (0.46)	0.73 (0.45)	0.63 (0.49)
Incremental cost per additional responder.	Dominated	21,090	-

sd = standard deviation

CBT and SC, as adjuncts to routine care, are both estimated to be more costly and more effective than RC. However, contrary to the study hypothesis, there was little difference between CBT and SC, and in fact SC is estimated to be more cost-effective than CBT, which it dominates. The ICER for SC as an adjunct to routine care, compared to RC, is estimated as £21,090 per additional responder.

3.1 Comparison of Methods 1 to 4

Four alternative methods for constructing CEACs from multiply imputed datasets are proposed in section 2. In this section, a brief illustration is given of how each method was applied to the SoCRATES trial dataset, and then the results of each method are compared.

Method 1: Bootstrapping the stacked datasets

The five datasets resulting from the multiple imputation were stacked to form one large dataset. 1000 bootstrapped samples of costs and health benefits were generated from this stacked dataset, from which to construct the CEACs.

Method 2: Bootstrapping each dataset separately

1000 bootstrapped samples of costs and health benefits were generated from each dataset separately. These were then combined using the correct formulae for calculating pooled means in order to provide 1000 samples from which to construct the CEACs.

Method 3: Simulating costs and outcomes

The five datasets were examined to inform the choice of distribution for the costs and health benefits. A lognormal distribution was chosen to describe the cost data, and this was informed with the pooled mean and pooled standard deviation from the multiply imputed datasets. We then recreated the original sample, 1000 times, and from each we calculated an estimate of the mean cost. In this trial we had a dichotomous outcome variable, equal to one if patients experienced a 20% improvement in PANSS total score at 18-months and zero otherwise. In order to directly simulate the sample, we would require a distribution, which returned values of zero or one in accordance with our pooled means and variations. Alternatively, a beta distribution can be used to describe the proportion of responders in each treatment group. The α and β parameters for the beta distribution can be calculated from the pooled mean and standard error using the method of moments formulae. Given that we know the sample size, the predicted average proportion of responders is sufficient to calculate the standard deviation within each sample, for comparison with the other methods.

Method 4: Imputing and simulating net monetary benefits

Costs were combined with health benefits to calculate net monetary benefits by applying appropriate values of WTP[19]. In total, 40 values of WTP were used to provide an adequate range over which to construct CEACs. Only then was the data imputed. The resulting datasets were examined to inform the choice of distribution for net benefits. The most suitable distribution was not obvious, which is unsurprising given that net monetary benefits in this example are calculated by combining lognormally distributed costs with a dichotomous outcome variable. When selecting a distribution, it is important to remember that the aim is not merely

to recreate the particular sample under analysis, but instead the aim is to describe the population from which that sample is drawn. Unfortunately, there is no convention as to which distributions are most suitable for describing net monetary benefits. A normal distribution was used for this analysis, informed by the pooled mean and pooled standard deviation. The properties of the normal distribution are suitable for describing net benefit, as it is not bounded at the upper or lower end of the range, and can be both positive and negative.

As is evident in the description of Methods 1 to 3, no attempt was made to impose a correlation structure on the data used to construct the CEACs. The main reason for this is that the appropriate structure was not known. In section 1.5 we asserted that a complete-case analysis was unlikely to be representative of the complete patient population. There is therefore no reason to suspect that the correlation structure observed in the complete-case analysis is any more suitable for informing the construction of these CEACs. Furthermore, the imputation procedure and the covariates chosen for imputation necessitated that costs and health benefits be imputed separately, and this may have introduced some spurious correlation. This does not detract from the fact that this is an issue that must be considered, and discussed, in a full cost-effectiveness analysis.

3.2 Results from Methods 1 to 3

Table 4 shows the within sample standard deviation in each of the 1000 bootstrapped or simulated samples, according to the method used. There is little difference between the methods in terms of the outcomes, which is in part due to the lower amount of missing outcome data. As such, we will focus on the different estimates of the variation in costs produced by each method. Contrary to our expectation, the within sample variation is not any larger in Method 3, despite Methods 1 and 2 ignoring the between-imputation component of the variance. We can also see that Method 2 allows us to estimate the correct within-sample standard deviation, but that calculating the standard deviation with the standard formula, rather than the correct formula for the pooled estimate, reflects a smaller amount of variation. Method 2 also serves to show that the between-imputation variance in this example is small. This is in part due to the largest cost component (inpatient resource use) being almost fully observed, and the majority of the imputed data being of much smaller magnitude (patient questionnaire data).

Table 4: Average within sample standard deviation in Method 1 to 3 compared to the pooled estimates from the multiply imputed datasets.

METHOD	Standard deviation in costs			Standard deviation in outcomes		
	CBT	**SC**	**RC**	**CBT**	**SC**	**RC**
Method 1	18762	17283	18061	0.45	0.44	0.48
Method 2						
- calculated using correct formula	18933	17266	18180	0.45	0.44	0.49
- ignoring between imputation variance	18844	17181	18076	0.45	0.44	0.49
Method 3	18389	17106	17904	0.45	0.44	0.48
Direct estimates from multiply imputed data (shown in Table 3)	18913	17412	18216	0.46	0.45	0.49

Table 5 illustrates the between sample standard deviation across the 1000 bootstrapped or simulated samples, according to the method used. It is this variation, which is ultimately reflected in the CEACs. Here, the variation is higher under Method 3, compared to Methods 1 and 2. However, there is an important caveat to be made. The bootstrapped samples are limited to the original observations whereas the observations in the simulated samples are free to take any value from the specified distribution. As such, we expect a higher amount of between sample standard deviation with the parametric methods. We would hope that some of the increased variance is due to the proper inclusion of the between-imputation variance, but there is no way to separate and quantify these two effects.

Table 5: Average between sample standard deviation in Method 1 to 3 compared to the pooled estimates from the multiply imputed datasets.

METHOD	Standard deviation in costs			Standard deviation in outcomes		
	CBT	**SC**	**RC**	**CBT**	**SC**	**RC**
Method 1	848.99	753.77	782.17	0.02	0.02	0.02
Method 2	557.71	508.98	534.77	0.01	0.01	0.01
Method 3	1791.45	1701.20	1819.52	0.05	0.04	0.06

3.3 CEAFs from Methods 1 to 4

Figures 1 to 4 show the CEAFs generated by each method. Method 4 is marked out as distinct from Methods 1 to 3 in the optimal treatment strategy at the upper end of the range of values for WTP. Using Method 4, CBT appears to be the most effective treatment strategy, rather than SC, but only at very high values of WTP (>£90,000). Method 4 employs the parametric method detailed in Method 3 to generate the CEAF, and also embeds the correlation between costs and outcomes in the both imputation,

and CEAF-generating procedures. However, as we mentioned earlier this method restricts the available cases used in the imputation procedure to only those with full cost and outcome data (25% of the full sample) and these may not be representative of the complete population. It is possible that by embedding the correct correlation structure in the imputation method, the inference was altered, but we must be cautious in our interpretation of this result.

The inference from Methods 1 to 3 does not change markedly with the method used to obtain the CEAF. In all cases routine care alone appears most likely to be cost-effective up to a WTP of about £20,000 per additional responder, whereafter SC is most likely to be cost-effective. The WTP value where the CEAFs for routine care alone and SC cross is, £21,317, £19,615, and £20,640 for Methods 1, 2 and 3 respectively. If the WTP for an additional responder is likely to be much less than these values, the method used will not impact upon the result. However, if the WTP was close to the point at which the optimum treatment changes, for example £20,000, the choice to bootstrap repeated samples or simulate repeated samples would affect the result. The choice of method also affects the level of certainty about the cost-effectiveness of each treatment at different values of WTP. Figures 5 to 7 display the CEACs generated by each method, separately by treatment. Fenwick *et al* (2001)[18] point out that the uncertainty characterised in CEACs and CEAFs can be used in calculating the expected value of information (VOI). The expected VOI involves estimating the value of obtaining more information on costs or outcomes in order to reduce the uncertainty associated with the treatment decision. The method used to generate the CEACs will clearly affect the calculation of VOI. The probability that each of the three alternative treatment strategies is cost-effective must always sum to one. If the between-imputation variance increases the uncertainty around the cost-effectiveness of the optimal treatment strategy, it must decrease the uncertainty around some sub-optimal strategies. If the between-imputation variance has a differential impact according to the treatment under consideration, then its omission will introduce a bias into the resulting CEACs.

4. Conclusions

In this paper, we have demonstrated that once multiple imputation has been used to address the problem of missing data, subsequent analyses should properly reflect the imputation procedure. Analysts should also be aware of the effect of the imputation

procedure on any correlation between costs and health benefits, and if possible this correlation should be included in subsequent analyses of the data. We have shown that a proportion of the variance associated with multiply imputed data cannot be reflected if standard, non-parametric methods are used to construct CEACs comparing more than two treatment options. It is difficult to quantify the effect of this omission, or any bias it may introduce. We have shown that it is possible for the quantitative results to be affected, as well as the implementation decision, depending on the threshold value of WTP, and as such parametric methods which properly incorporate the between-imputation variance are likely to be more suitable for analysing multiply imputed datasets.

Bayesian methods, which can handle the issues of missing data, correlation between costs and outcomes, and the analysis of cost-effectiveness simultaneously are likely to be the way forward in addressing the issues considered here.

Figure 1: CEAF generated by Method 1

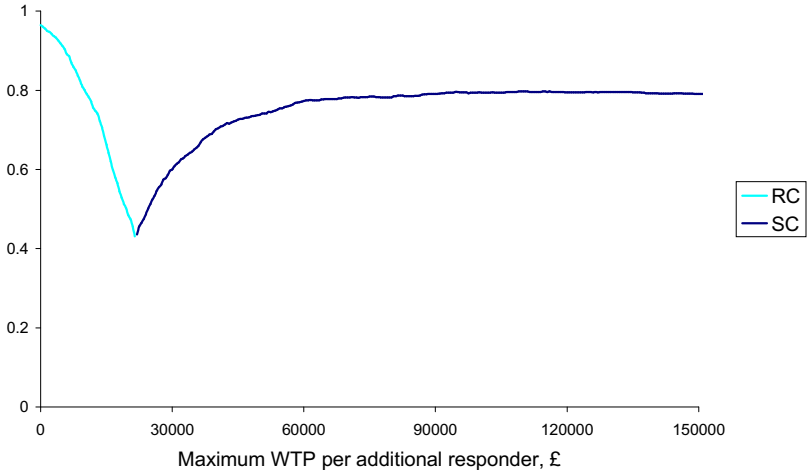


Figure 2: CEAF generated by Method 2

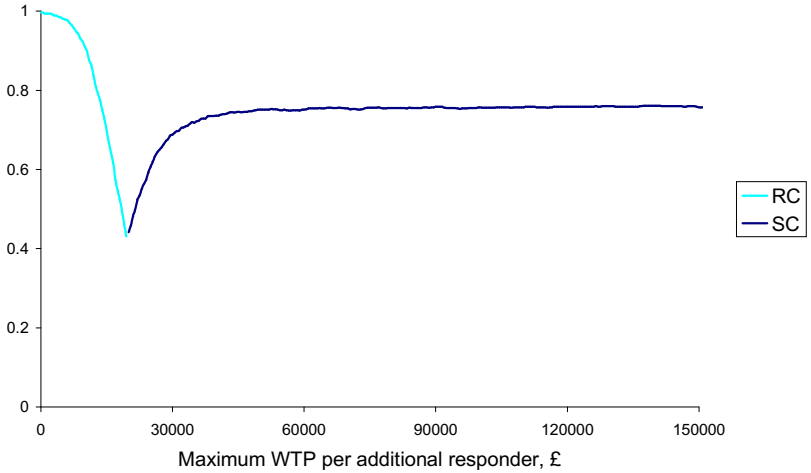


Figure 3: CEAF generated by Method 3

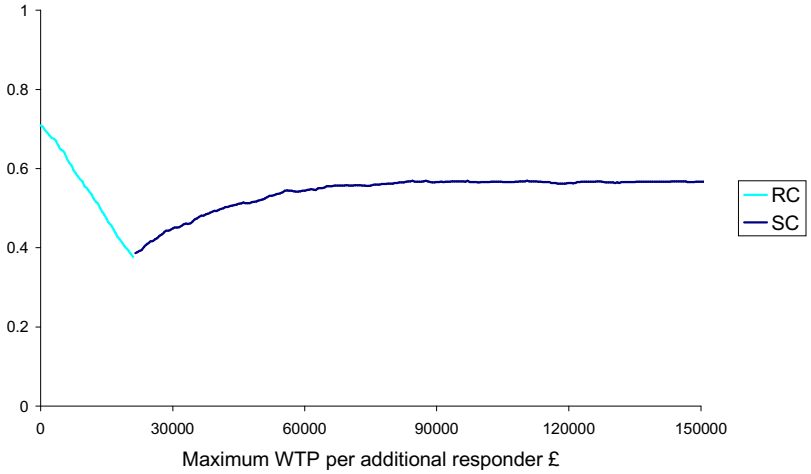


Figure 4: CEAF generated by Method 4

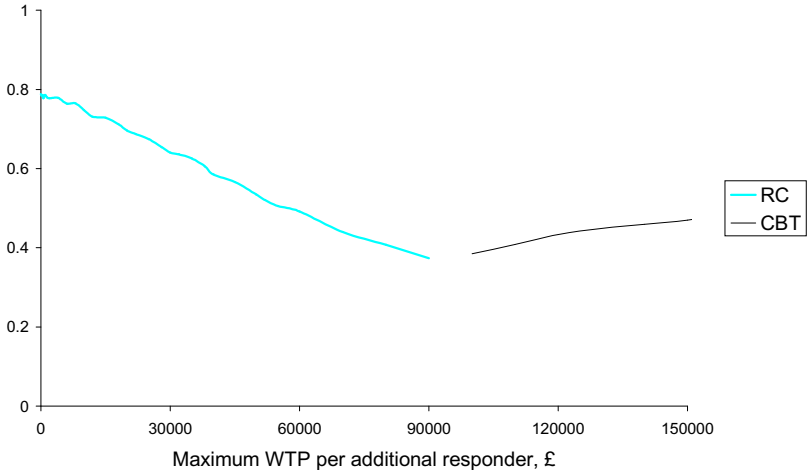


Figure 5: CEACs for CBT generated by Methods 1 to 4

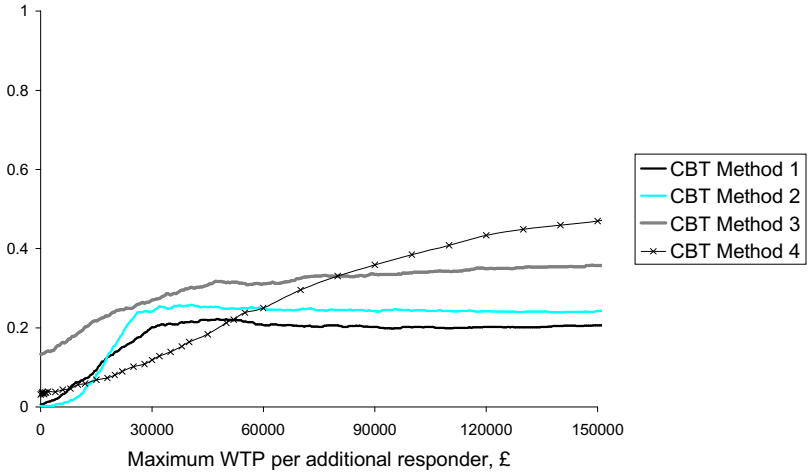


Figure 6: CEACs for SC generated by Methods 1 to 4

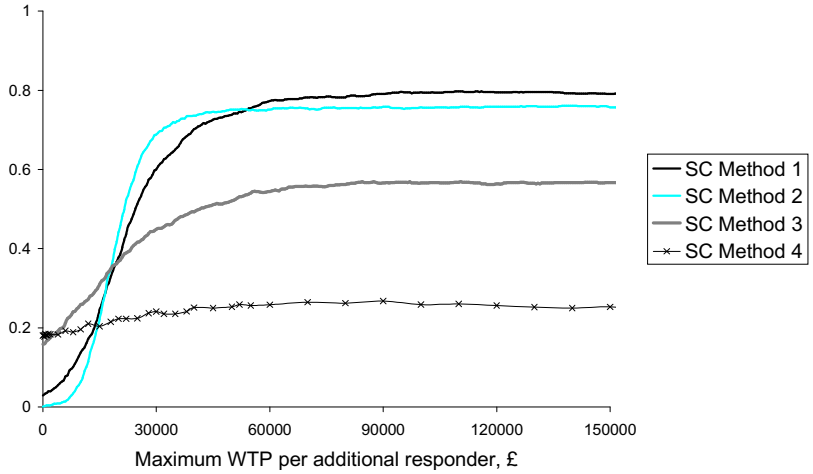
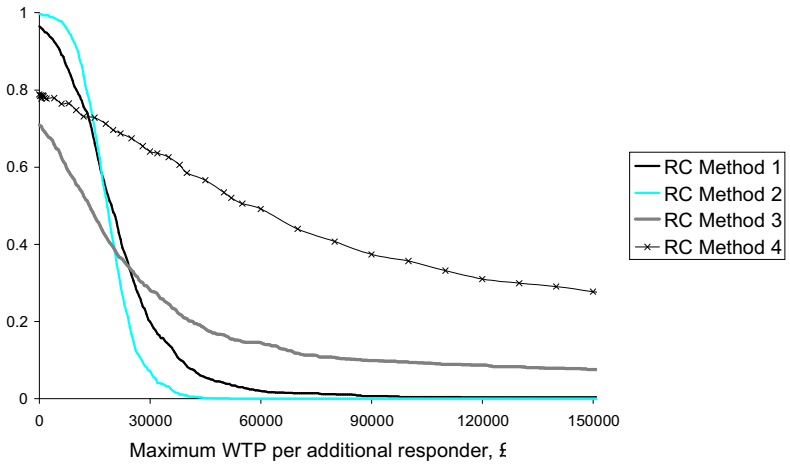


Figure 7: CEACs for RC generated by Methods 1 to 4



Acknowledgements
I would like to acknowledge the input of Stephen Palmer, my supervisor, and Andrea Manca at the Centre for Health Economics, University of York.

References

1. Briggs, A., et al., *Missing.... presumed at random: cost-analysis of incomplete data.* Health Economics, 2003. **12**(5): p. 377-392.
2. Oostenbrink, J.B., J.A. Maiwenn, and M.P.M.H. Rutten-van Mölken, *Methods to Analyse Cost Data of Patients Who Withdraw in a Clinical Trial Setting.* Pharmacoeconomics, 2003. **21**(15): p. 1103-1112.
3. Manca, A. and S. Palmer. *Handling missing data in economic evaluation alongside randomised clinical trials.* in *HESG*. 2001.
4. Scott, J., et al., *Use of cognitive therapy for relapse prevention in chronic depression.* British Journal of Psychiatry, 2003. **182**: p. 221-227.
5. Lewis, S., et al., *Randomised controlled trial of cognitive-behavioural therapy in early schizophrenia: acute-phase outcomes.* British Journal of Psychiatry, 2002. **181**(supplement 43): p. 91-97.
6. Kay, S.R., L.A. Opler, and J.-P. Lindenmayer, *The Positive and Negative Symptom Scale (PANSS): rationale and standardisation.* British Journal of Psychiatry, 1989. **155**(supplement 7): p. 49-52.
7. Gureje, O., et al., *Olanzapine vs risperidone in the management of schizophrenia: a randomized double-blind trial in Australia and New Zealand.* Schizophrenia Research, 2003. **61**(2-3): p. 303-314.
8. Dossenbach, M.R., et al., *Evidence for the effectiveness of olanzapine among patients nonresponsive and/or intolerant to risperidone.* Journal of Clinical Psychiatry, 2001. **62**(supplement 2): p. 28-34.
9. Lindenmayer, J., et al., *Olanzapine in refractory schizophrenia after failure of typical or atypical antipsychotic treatment: an open-label switch study.* Journal of Clinical Psychiatry, 2002. **63**(10): p. 931-935.
10. Spina, E., et al., *Relationship between plasma risperidone and 9-hydroxyrisperidone concentrations and clinical response in patients with schizophrenia.* Psychopharmacology, 2001. **153**(2): p. 238-243.
11. Knapp, M. and J. Beecham, *Costing mental health services: The Client Service Receipt Inventory.* Psychological Medicine, 1990. **20**: p. 893-908.
12. Netten, A. and L. Curtis, *Unit Costs of Health and Social Care 2001.* 2001, University of Kent at Canterbury: Personal Social Services Research Unit.
13. Association, B.M. and R.P.S.o.G. Britain, *British National Formulary*. 2002.
14. *SOLAS for missing data analysis*. 2001, Statistical Solutions Ltd.: Cork.
15. Rubin, D.B., *Multiple Imputation for Nonresponse in Surveys.* 1987, New York: John Wiley and Sons.
16. Johannesson, M. and S. Weinstein, *On the decision rules of cost-effectiveness analysis.* Journal of Health Economics, 1993. **12**: p. 459-467.
17. Van Hout, B.A., et al., *Costs, effects and c/e-ratios alongside a clinical trial.* Health Economics, 1994. **3**: p. 309-319.
18. Fenwick, E., K. Claxton, and M. Sculpher, *Representing uncertainty: the role of cost-effectiveness acceptability curves.* Health Economics, 2001. **10**: p. 779-789.
19. Stinnett, A.A. and J. Mullahy, *Net Health Benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis.* Medical Decision Making, 1998. **18**(supplement 2): p. 68-80.
20. *Crystal Ball 2000*, Decisioneering, Inc.: Denver.
21. *Excel*, Microsoft Office 2000.