Estimating cost-effectiveness from clinical trials: Problems with hypothesis testing and missing data

Mirella Longo, Wai-yee Cheung, Andy Briggs, Kerenza Hood.

University of Glamorgan, Business School
University of Wales, Swansea, School of Postgraduate Studies in Health and Medical Care
University of Oxford, Health Economics Research Centre
University of Wales College of Medicine, Cardiff, Department of General Practice.

Introduction

RCTs are commonly accepted as the gold standard for evaluating clinical effectiveness, and economic evaluations alongside RCTs are increasingly employed to address issues of cost-effectiveness (Sculpher *et al,*1997). There are several advantages in performing economic evaluation alongside RCTs (Drummond *et al*, 1997), and the possibility of using stochastic data certainly is one of the most important. Moreover, in a regime of limited resources, policy makers have become increasingly interested of the resource implication of new interventions, and economic alongside RCTs seems the best vehicle to cast this information.

To date, although much more research is still needed, the number of statistical issues investigated on the economic evaluation alongside RCTs is constantly increasing, and they supply researchers with sophisticated and precise techniques to improve estimation and confidence intervals calculations (Chaudhary *et al,*1996, Briggs *et al,*1997,).

In this paper, we would like to take a step back and consider two issues in economic evaluation alongside RCTs. The first concerns the choice between comparison of parametric and non-parametric analyses, the second deals with the analysis and imputation of missing data. These two issues are paramount in any statistical analysis as, if overlooked, they could compromise the validity of any further analysis and generate misleading conclusions.

As the content of both issues is very large, we will focus on the comparison between t-test and Mann Whitney U test for the first issue and on the use of multiple imputation approach for data missing at random and data missing not at random for the second issue. Two case studies are used for illustration.

T-test and Mann-Whitney U test

The word non-parametric evolves from the type of hypothesis usually tested when dealing with ordinal data, and most non-parametric tests do not involve inferences about parameters from the original population. Instead of hypothesising that the two

populations have the same mean (t-test), we could hypothesise that the two populations from which the two samples are drawn are identical (Mann Whitney U test). The implications of these two hypotheses are quite different. In the first case the null hypothesis is specific to a particular population parameter, whereas in the latter case the null hypothesis addresses the question of equality of the probability distribution. Two distributions could be different and still have the same mean (see study case 1).

Also, because ordinal or categorical data and category frequencies do not require the existence of a meaningful numerical scale for measurements, the null distribution of a non-parametric statistic can be determined without regard to the shape of the underlying population distribution. For this reason, these procedures are also called distribution free-tests. This distribution free property is their strongest advantage, and many statisticians refer to these as distribution-free analyses rather than non-parametric analyses (Bhattacharyya GK and Johnson RA, 1977).

To this point two conclusions seem to be clear:

1) When dealing with ordinal/categorical data, because the distance between values does not lead to any practical interpretation, non-parametric procedures that utilise information only on order/category/rank are preferred to parametric statistics.
2) When dealing with continuos data, and the population is known to be normally distributed, the parametric procedures are certainly more efficient.

However, although, non-parametric statistical tests have been developed specifically for ordinal data, they are also appropriate (indeed they are recommended) for quantitative data when one or more of the assumption underlying a particular statistical test has been violated (Greenhalgh T, 1997, Bland M, 1997).

Dealing with missing data

It is important to stress out that an economic evaluation is only as good as the data upon which it is based, but economic analysts must do the best they can with the data available (Drummond *et al*, 1997). The collection of data alongside RCTs raises two important issues: the need for a balance between comprehensiveness (i.e. collection and analysis of resource use and outcome) and manageability in terms of data handling and the resource implication of prospective data collection (Coyle *et al*, 1994).

Data on cost and effectiveness alongside RCTs usually implies the use of several data collection methods (i.e. patient self-report, medical records). Missing data are an inevitable feature of any data collection process and, the researcher's challenge is to address the missing data problem to improving the generalizability of the results.

When performing missing data analysis, the researcher's primary concern is to determine the reasons underlying the missing data (the missing data mechanism). The pattern of missing data can assume three forms (Little and Rubin, 1987):

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.

Dichotomised correlations and scatter plots are the most popular methods used to diagnose the missing data process. In the former, valid values are represented by the value of one, and missing data are replaced by the value of zero. These missing values for each value are then correlated; the correlation indicating the degree of association between the missing data of each variable pair (Hair *et al,*1984).

If data are either MCAR or MAR the methods suggested to deal with missing data are:

1) Complete-case analysis
2) Available case analysis
3) Mean Imputation
4) Hot-decking
5) Last value carried forward
6) Regression imputation
7) Maximum likelihood imputation using the EM algorithm

These techniques are amply illustrated elsewhere (Hair *et al,*1984, Wolstenholme J and Briggs A, 1999).

Although CCA and ACA present several weaknesses they completely rely on real data, and this is to some extent the main critic towards the methods 3 to 7. In these cases, the estimation process underestimates the SEs and variances of the mean and regression coefficients. In order to improve variance estimate, total variance must include a term that accounts for the amount of uncertainty involved in the imputation of the missing values. A number of imputations should be performed.

Multiple imputation is performed by generating m (say 5) data files and, for each data set, performing the desired statistics: estimate $\hat{q}i$ and its estimated variance $\hat{\text{var}}(\hat{q}i)$. The results are then combined and the imputation of q is $\hat{q} = \frac{1}{M}\sum_{i=1}^{M}\hat{q}i$ and its estimated variance is

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

= mean of estimated variance $\hat{\text{var}}(\hat{q})$ + variance of the estimates $\hat{q}i$.
The fist term on the left relates to the variance within the imputed data sets and the term on the right hand side captures the uncertainty due to the variability in the imputed values (The term 1 = 1/M is a bias correction factor) (Wolstenholme J and Briggs A, 1999).

The two case studies differ in the sense that while the first has data missing at random, the second presents data not missing at random. Thus, before multiple imputation could be performed, alternative ways to estimate the missing data have been used.

Case Study 1: Long Term Follow-up of Patients with Irritable Bowel Disease: Open Access v. Routine Appointments.

Background:

This study was a randomised trial assessing access follow up of patients with inflammatory bowel disease (the intervention) as compared with current practice of follow up via routine booked appointments (control). 180 patients were recruited and followed up for 24 months.

The primary outcome measure was a condition specific questionnaire, the UK-IBDQ, which combines the scores of its various sub-scales into a single index. To counter for possible differences at baseline, changes in scores between baseline assessment and 24 month follow up were used. Costs to the NHS in both primary and secondary care were assessed. Resource use data in primary care were collected through postal questionnaires and semi-structured interviews and in secondary care through reviewing hospital notes.

Issue 1: parametric v. non-parametric testing.

Table 1 below shows the key variables of the economic evaluation. As can be seen, the two approaches lead to a different conclusion for total secondary care costs (Mann-Whitney $p<.05$, t-test p = non-significant).

Table 1. Case study 1 - costs and outcomes items.

Vairable	Study Group	N	Mean (SD)	Mean Rank	p.value*	p.value**
Outcome measures at recruitment	Intervention Control	87 89	72 (16) 70 (19)	90 87	ns	n/a
Outcome measures at 24 months	Intervention Control	80 81	69 (19) 66 (19)	85 77	ns	n/a
£Out-patient visit	Intervention Control	88 91	251(205) 297(150)	76 104	ns	<.0001
£In-patient stay	Intervention Control	88 91	177 (677) 117 (429)	91 89	ns	ns
£Tests and Investigations	Intervention Control	88 90	205 (298) 263 (275)	81 98	ns	<.05
£Tot NHS Secondary Care	Intervention Control	88 90	634 (903) 670 (589)	79 100	ns	<.05
£Total Drugs	Intervention Control	76 79	378 (467) 263 (401)	83 74	ns	ns
£GPs Surgery consultation	Intervention Control	76 79	65 (55) 61 (50)	79 78	ns	ns
£Home visits	Intervention Control	76 79	19 (59) 21 (62)	77 79	ns	ns
£Tot NHS Primary Care	Intervention Control	76 79	462 (472) 344 (427)	84 73	ns	ns
£Tot NHS Primary and Secondary care	Intervention Control	76 78	1032 (959) 961 (678)	76 79	ns	ns

* T-test
** Mann-Whitney U-test

Returning to the definition of Mann-Whitney U test, the null hypothesis tests if the two population distributions are identical. What we have demonstrated in our case is that the two population distributions for secondary care costs are different. In fact, although we have a difference of almost £120 for primary care costs, this does not

reach statistical significance by either method. In this case, the distribution of the two population curves (location of the data) is very similar and the mean rank is close.

The estimates of differences generated from the Mann-Whitney U test would give the difference between the mean ranks but, mean ranks unable us to determine C/E ratios. Hence, the Mann-Whitney U test is better suited to hypothesis testing than to estimation (Briggs A and Gray A, 1998).

It is also worth making a further point. From basic descriptive statistics (i.e. histogram) we could determine with confidence that the health outcomes values are normally distributed. The same does not apply to the resource use data. However, as sample size increases the distribution of the sample statistics (i.e. mean, SD) tend to be normal. This even if the population does not follow a normal pattern (central limit theorem). The skew coefficient of the population will be reduced by a factor of $\sqrt{n}$ in the sampling distribution of the mean of the population, where n is the sample size. This is directly derived from the definition of skew coefficient (Kendall MG and Stuart A, 1969). Some suggest 50 as the minimum sample size needed for the central limit theorem to apply (Bland, 1997).

If the data are in fact highly skewed, then alternative approaches to the Mann-Whitney U test must be found. One way forward is to log transform the data, this would be particularly suitable when we have long tails as in resource use. However, because the log transformation compares geometric means, this would unable us to determine the confidence intervals of the mean difference (Briggs A and Gray A, 1998) and ultimately the C/E ratios.

An alternative approach similar to the t-test but that allows us to relax the assumption of normality of the distribution is the non-parametric bootstrap. The idea here is that, in the absence of any further knowledge on the population, the distribution of values in a sample of size n randomly drawn from that population, is the best guide to the distribution in the population and repeated re-sampling can provide an approximation of the population distribution (Mainly BFJ, 1997). One of the main themes in research on bootstrapping applications in economic evaluation alongside RCTs is to calculate valid confidence limits around the comparison between intervention and control groups (Chaudary MA *et al,* 1996, Briggs A and Fenn P, 1998, Obenchain RL *et al*, 1997). However, as this is not the aim of this paper we refer to the literature available for more details.

Issue no. 2 : Handling missing data

UK-IBDQ scores were obtained for 98% (176/180) of the study patients at recruitment, and 89% (161/180) at the end of the study. Differences between patient score at recruitment and score at the end of the study could be calculated for 89% (160/180) of the patients. Primary care cost were available for 86%(155/180) and secondary care cost for 99% (178/180). There were 140 sets of complete cases (78%).

Data was not, however, missing completely at random. Omitting cases with incomplete data would result in eight sets of cases with extreme high values (Q3 +

1.5*IQR) on costs (one in primary care and seven on secondary care cost) being excluded from the analysis. This could cause underestimate of the mean and the variance of the costs, especially for cost in secondary care as shown in Table 2 below.

Table 2: Mean (sd) outcome and costs: CCA v. ACA

	Complete Case Analysis (CCA) mean (sd)	Available Case Analysis (ACA) mean (sd)
IBDQ at recruitment	70.98 (17.75)	70.94(17.56)
IBDQ at 24 months	66.94(19.17)	67.43 (19.09)
NHS primary care cost	404.48 (426.60)	402.00 (452.39)
NHS secondary care cost	602.60 (675.61)	652.05 (658.81)

Examination of the pattern of missing values of the current data set showed very small correlations between the four cost-effectiveness variables and the other variables that could potentially be used as predictors for generating missing values (e.g. age, sex, diagnostic groups, study groups and study centres). The correlations between the four cost-effectiveness variables were small (r less than 0.2) except for the correlation between IBDQ at recruitment and at the end of study (r=0.608).

The standard variable by variable analysis (available case analysis) meant that the conclusions would be drawn from a divergent set of cases. Furthermore, standard multivariate methods are designed only for analysis of complete cases. Older methods of handling missing data by single imputation had been shown to be systematically under-estimating the variance of the actual values (Rubin DB, 1996). Therefore, multiple sets of values were generated for multiple imputation to solve the problem of obtaining valid statistical inferences.

As the correlations between the variables were small, missing values were generated by non-parametric bootstrap. Bootstrapping was performed on two source files, one containing all available cases (MI-A) and the other contained the complete cases (MI-C). For each source file, five sets of values were generated for the missing data points. To prevent diluting any possible differences between study groups, values were bootstrapped from within the appropriate study group.

While bootstrapping from available cases, replacement values for the missing data items were on costs and effectiveness were bootstrapped separately. There was high correlation between IBQ scores at recruitment and follow up. Therefore, replacement values for the 20 missing data points on the outcome measure were generated from the 160 cases available. Replacement values for the missing data points on costs were generated from 155 cases for primary care costs data, 178 cases for secondary costs data and 154 cases for total NHS cost data. This approach makes maximum use of the observations but it might run the danger of distorting the relationship between resource use and effectiveness, and possible relationship between primary and secondary care costs.

When bootstrapping form complete cases, replacement values for missing data on all the variables were bootstrapped together. Replacement values for all variables of cases with any incomplete data were generated from the 140 cases. This approach

preserves possible relationships between the variables but entails omitting cases with incomplete data from the final analysis. We compared this two approaches with ACA and CCA without imputation to identify possible differences in conclusions. Results showing the different methods of analysis are shown in Tables 3-6.

Table 3: Differences in UK - IBDQ scores using 4 methods:
Available Case Analysis (ACA), Complete Case Analysis (CCA), bootstrap Multiple Imputation of Available cases (MI-A) and Complete cases (MI-C)

	Open Access				**Routine**			
	ACA	CCA	MI-A	MI-C	ACA	CCA	MI-A	MI-C
n	79	69	88	88	81	71	92	92
mean	-3.79	-4.76	-3.68	-4.9	-3.42	-3.35	-3.55	-3.3
sd	18.21	18.85	18.1	19.26	13.19	13.68	13.07	13.39
diff	-0.37	-1.4	-0.12	-1.6				
SEdiff	13.35	13.87	2.57	2.68				
t-ratio	-5.03	-0.5	-0.05	-0.6				
p-value	0.881	0.618	0.952	0.552				
95% CIs	-5.35 to 4.6	-6.92 to 4.11	-5.16 to 4.91	-6.85 to 3.66				

Table 4: Primary care costs using 4 methods:
Available Case Analysis (ACA), Complete Case Analysis (CCA), bootstrap Multiple Imputation Available cases (MI-A) and Complete cases (MI-C)

	Open Access				**Routine**			
	ACA	CCA	MI-A	MI-C	ACA	CCA	MI-A	MI-C
n	76	69	88	88	79	71	92	92
mean	462.11	479.07	475.65	479.84	344.17	331.99	366.14	336.11
sd	472.38	488.43	473.69	485.29	427.32	344.7	459.27	348.72
diff	117.94	147.08	109.5	143.73				
SEdiff	72.44	71.63	77.6	64.04				
t-ratio	1.63	2.05	1.41	2.24				
p-value	0.106	0.042	0.161	0.025				
95%CIs	-25.20 to 261.08	5.46 to 288.69	-42.59 to 261.59	18.2 to 269.25				

Table 5: Secondary care costs using 4 methods:
Available Case Analysis (ACA), Complete Case Analysis (CCA), bootstrap Multiple Imputation Available cases (MI-A) and Complete cases (MIC)

	Open Access				**Routine**			
	ACA	CCA	MI-A	MI-C	ACA	CCA	MI-A	MI-C
n	88	69	88	88	90	71	92	92
mean	634.02	587.93	634.02	582.72	669.68	616.86	670.80	610.5
sd	903.32	831.9	903.32	814.48	588.92	483.4	584.89	490.48
diff	-35.66	-28.92	-36.78	-27.78				
SEdiff	114.57	115.42	114.27	120.47				
t-ratio	-0.31	-0.25	-0.32	-0.23				
p-value	0.757	0.803	0.749	0.819				
95% CIs	-261.82 to 190.5	-257.10 to 199.26	-260.75 to 187.18	-263.9 to 208.33				

Table6: Total NHS costs using 4 methods
Available Case Analysis (ACA), Complete Case Analysis (CCA), bootstrap Multiple Imputation Available cases (MI-A) and Complete cases (MIC)

	Open Access				**Routine**			
	ACA	CCA	MI-A	MI-C	ACA	CCA	MI-A	MI-C
n	76	69	88	88	78	71	92	92
mean	1031.67	1067	1046.69	1062.56	961.03	948.85	955.47	956.61
sd	903.32	973.52	972.48	957.59	678.29	618.99	672.68	623.12
diff	7.64	118.15	91.22	115.95				
SEdiff	134.16	138.32	188.47	200.74				
t-ratio	0.53	0.85	0.48	0.58				
p-value	0.6	0.95	0.64	0.58				
95% CIs	-194.45 to 335.74	-155.30 to 391.61	-278.19 to 460.62	-277.5 to 509.39				

For the current data set, the point estimates and the width of the interval (95% CI) estimates of the differences reached by available case analysis, complete case

analysis, multiple imputation from available case, multiple imputation from complete case are similar in all parameters.

However, if we move from estimation to hypothesis testing, the four sets reach different conclusions for one of the parameter. The primary care costs did not reach statistical significance when CCA and ACA analyses were performed while statistical significance was reached by imputation of missing data from complete case available and available cases.

Case study No. 2. Guidelines for the Referral and Management of Women with Breast Disorders

Background:

In 1996, the Department of Health developed guidelines regarding the primary care management of common breast presentations. The principle aim was to ensure appropriate referral of women with symptoms suggesting a high likelihood of cancer, and promote community management of women with symptoms suggesting a low likelihood of cancer. Separate guidelines were produced for women with presenting symptom of 1) pain 2) lumpiness and 3) discrete lump.

There is, however, evidence that simply posting guidelines go general practitioners does not have a major effect on practice (Grimshaw JM and Russell IT, 1993). A randomised study was therefore undertaken to assess to what extent an educational intervention could alter the clinical management of patients by GP's.

Following a baseline data collection phase, general practices were randomised to receive education and training in either the pain guidelines alone (and thus act as controls for the lump/lumpiness practices) or in the combined lump/lumpiness guidelines alone (and thus act as controls for the pain practices). Given the principle aim of the guidelines, the primary outcome measure of the trial (the unit of effectiveness) was "concordance of patient management with guidelines". A cost effectiveness analysis was run alongside the trial.

Thirty four GP practices were involved in the study, 16 of which were randomised to receive training in the pain guidelines and 18 of which were randomised to receive training in the lump/lumpiness guidelines. A total of 944 women were recruited to the study, 410 of whom were patients in 'pain practices' and 534 of whom were patients in 'lump/lumpiness practices'. Thus a woman presenting with breast pain was thus in the intervention group if she belonged to a 'pain practice' and in the control group if she belonged to a 'lump/lumpiness practice' and vice versa.

<u>Issue No. 1: Parametric v. non-parametric testing</u>

The data from this study were in the main not highly skewed. Table 7 below shows primary and secondary care costs for intervention (i.e. women with pain in pain practices, women with lump/lumpiness in lump/lumpiness practices) and control (i.e. women with pain in lump/lumpiness practices, women with lump/lumpiness in pain

practices) groups. Here, the non-parametric Mann-Whitney U test and the parametric t-test both show a statistically significant difference in drug costs for women with pain and no significant differences across any other resource use variables. Thus, unlike the situation in case study no. 1, the small degree of skewness of the resource use data in case study no. 2, does not produce conflicting messages whichever test is used.

Table 1. Case study 2 – items of primary care cost

Items	Study Group	N	Mean (SD)	Mean Rank	p.value*	p.value**
Symptom of Lump						
£GP follow up visit	Intervention Control	190 188	41 (35) 43 (35)	185 193	ns	ns
£Drugs	Intervention Control	202 214	1(4) 0.65(3.44)	211 206	ns	ns
£Total NHS Primary Care (GP visits, Drugs, Training)	Intervention Control	178 183	48(35) 50 (35)	177 185	ns	ns
Symptom of Lumpiness						
£GP follow up visit	Intervention Control	81 76	49 (47) 46 (35)	80 78	ns	ns
£Drugs	Intervention Control	82 94	6 (10) 5 (10)	92 85	ns	ns
£Total NHS Primary Care (GP visits, Drugs, Training)	Intervention Control	77 76	63 (43) 57 (36)	79 75	ns	ns
Symptom of Pain						
£GP follow up visit	Intervention Control	167 107	48 (36) 45 (41)	142 130	ns	
£Drugs	Intervention Control	208 109	4 (7) 9 (11)	148 179	<.0001	<.001
£Total NHS Primary Care (GP visits, Drugs, Training)	Intervention Control	164 107	59 (36) 60 (43)	137 134	ns	ns

*T-test

** Mann-Whitney U-test

Issue No. 2: Missing Data

The tables 8-10 describe the variables investigated together with the data that was available and missing. To improve presentation, the data are presented according to the data collection source. Data on primary care and drug use were retrieved from GPs notes, secondary care data were collected from hospital notes.

Primary care data

Because the percentage of missing data for drugs is relatively small the missing values can be confidently estimated by mean substitution. The missing data on number of visits to the surgery is already to big to consider to use mean substitution but, as data are missing at random, the re-sampling method applied for case one can be applied.

Table8: Data available and missing data retrieved from GP notes

Item	N	Missing	
		Count	Percent
Patient age	923	21	2.2
WTE (All time equivalent GP)	944	0	0
GP visits to the surgery	809	135	14.3
NSAID	909	35	3.7
Gamolenic	910	34	3.6
Danazol	910	34	3.6
Bromocriptine	910	34	3.6
Other type of drugs	891	53	5.6
Where was the patient referred	886	58	6.1
County (Geographical location of the Practice)	944	0	0
GP gender	914	30	3.2
Fund Holding status of the practice	944	0	0
Presenting symptom	944	0	0
Presenting sign	944	0	0

Secondary care data

315 of our cohort of women were not referred to a consultant, as result these are not applicable cases for the secondary care missing data analysis and they have to be taken out. This leaves a cohort of 629 women referred to secondary care.

Table 9 lists data available and missing data for each item collected via hospital notes. The main reason for the large number of missing data is that two of seven centres supplied us with data in a format that made it impossible to link the observations to the patients. The only information that can be gleaned from the data as they stand is the total number of cases per centre (i.e. total number of wide bore biopsy). This can, however, be of some use in that it can help to validate our estimations.

Table9: Data available and missing data retrieved from hospital notes

Item	N	Missing	
		Count	Percent
Assessment (Consultant visit)	366	263	41.8
Mammography	365	264	42
Breast Ultrasound Scan	365	264	42
Breast Fine Niddle	365	264	42
Breast Diagnostic Biopsy	365	264	42
Breast Wide Bore Biopsy	365	264	42
Breast Cyst Aspiration	365	264	42
Breast X-ray	365	264	42
Bone scan	365	264	42
USS Liver	365	264	42
Diagnosis Left Breast	347	282	44.8
Diagnosis Right Breast	351	278	44.2
Management: Follow up	347	282	44.8
Management: Breast cancer	347	282	44.8
Management: Family history cancer	346	283	45
Management: Mastalgia	346	283	45
Exam (Who did exam the woman)	356	273	43.4
Consultant (Did a consultant examine the woman?)	366	263	41.8
Symptom still present when patient arrived to the clinic	351	278	44.2
Right breast lump detected on examination	354	275	43.7
Right breast lumpiness detected	354	275	43.7
Right breast other symptoms detected	354	275	43.7
Left breast lump detected on examination	354	275	43.7
Left breast lumpiness detected on examination	354	275	43.7
Left breast other symptoms detected	354	275	43.7
Grading score right breast: Clinical	196	433	68.8
Grading score right breast: Mammography	214	415	66
Grading score right breast: Ultrasound	287	342	54.4
Grading score right breast: Imaging	205	424	67.4
Grading score right breast: Cytology	287	342	54.4
Grading score right breast: Histology	336	293	46.6
Grading score left breast: Clinical	206	423	67.2
Grading score left breast: Mammography	216	413	35.7
Grading score left breast: Ultrasound	298	331	52.6
Grading score left breast: Imaging	200	429	68.2
Grading score left breast: Cytology	304	325	51.7
Grading score left breast: Histology	335	294	46.7
Family history assessment (Low/High risk)	259	370	58.8
Family history assessment: First degree relatives	354	275	43.7
Family history assessment: Second degree relatives	354	275	43.7

In this example the data are clearly not missing at random and alternative ways of estimating the resource use in these two centres must be found. Moreover, if the complete case approach were adopted, something like 43% of the information would be lost, giving very poor confidence in the results (see table10).

Table 10: Data available and missing data for the resource use estimation

Item	N	Missing	
		Count	Percent
£GP Visits	809	135	14.3
£Drugs	909	35	3.7
£Secondary Care	680	264	28
£Total NHS	543	401	42.5

Imputation of data from the two centres

Because complete data sets are missing from two centres, and management between centres might vary. One way forward is to examine if a model can be developed which can predict secondary care data resource use and then examine if the model fit (R square) changes when the factor "centre" is included. If the two do not change, then we can assume that there is not so much of difference between the centres and we can use our model to predict the data missing from the two centres.

We used the variables available from primary care to predict secondary care costs, categorical variables were transformed in dummy variables first. As pointed above 315 women were not referred to secondary care, as result, the total secondary care variable would include 315 zero values. In order to reduce skewness these cases were excluded in the analysis. First results are shown in tables 11-13.

Table 11: Regression to predict secondary care costs

First regression: Method = Enter DV = Secondary care cost			Ivs: Age Population density WTE Total Primary care costs Family history Symptom Sign Site/type/severity of pain		
Anova					
Model	Sum of Square	df	Mean Square	F	Sig.
Regression	916595.03	18	50921.9	2.2	.002
Residuals	5917696.7	268	22080.9		
Total	6834291.7	286			

R square = .134
Adjusted R square = .076
SE of the estimate = 148.6

Table 12: Regression to predict secondary care costs adding centre as DV

Second regression: Method = Enter DV = Secondary care cost			Ivs: Age Population density WTE Total primary care cost Family history Symptom Sign Site/type/severity of pain Centre		
Anova					
Model	Sum of Square	df	Mean Square	F	Sig.
Regression	1004567.9	20	50228.4	2.2	.002
Residuals	5803126.6	257	22603.6		
Total	6813694.5	277			

R square = .147
Adjusted R square = .081
SE of the estimate = 150.3

Table 13: Regression to predict primary care costs

First regression: Method = Enter DV = Primary care cost			Ivs: Age Population density WTE Family history Symptom Sign Site/type/severity of pain Fund holding status GP gender		
Anova					
Model	Sum of Square	df	Mean Square	F	Sig.
Regression	91714.3	20	4585.7	3.4	<.0001
Residuals	1007500.7	750	1343.3		
Total	1099215.0	770			

R square = .083
Adjusted R square = .059
SE of the estimate = 36.65

Although the regression explains a relatively small proportion of the variance, the multiple imputation approach should correct for it. The results of the multiple imputation will be available on the web as soon as ready.

Discussion

Through the analysis of two cases studies we explored how the choice between hypothesis testing and estimation approach and the analysis of missing data might lead to divergent conclusions. Although, these issues can hardly be considered new research questions in statistics, the paper was mostly prompt from the fact that hardly any of the publications include description of missing data and Mann-Whitney U test is sometimes used for comparison on resource use.

Missing value description and analysis should be accepted as a necessary prerequisite for any publication. Policy makers need to make the best decisions they can today, recognizing that data are not (nor ever will be) perfect (Drummond M *et al,* 1997).

When dealing with skewed data one problem is related to the difficulties in separating “skewed data” from “highly skewed data”. Although it is realistic to accept the idea that no clear dividing line can ever be defined between "skewed" and "highly skewed" data, it is also true that most statistical textbooks do not offer any guidance that would help the researchers to take up a position they can defend.

Finally, although is generally known that RCTs are mainly designed on the basis of clinical effectiveness data (Claxton K *et al,* 1996, Al MJ *et al,* 1998). Reality is that economic as clinical data rarely come in the form and quantity we expected. This offers a challenge to researchers who, basically, have to reach the best estimate and define its likelihood on the data made available alongside the trial.

Further research:

In the case study one, omitting cases with incomplete data would result in eight sets of cases with extreme high values (Q3 = 1.5*IQR) on costs (One in primary and 7 in secondary care cost) being excluded from the analysis. It might be worth to investigate whether this is still a case of missing at random or not (Did we miss the more poorly patients?)

One decision to be made when performing missing data analysis is to decide whether to impute the missing data regardless of the arm of the trial or separate intervention from control group. The latter will have the effect of amplifying possible differences between intervention and control group. We felt this is less of a problem for case study one because it has been formulated as an equivalent trial. Possible difference between study group should be given the "benefit of doubt", This can be a problem for other types of trial.

In case study one we used a derivation of hot-decking approach to generate replacement values. We will try the regression approach and compare the results.

References:

Al Mj, Van Hout BA, Michel BC, Rutten FFH, Sample size calculation in economic evaluations, Health Economics, 1998;7:327-335.

Bhattacharyya GK and Johnson RA, Statistics Concepts and Methods, Wiley series in probability and mathematical statistics, 1977.

Bland M. An Introduction to Medical Statistics, Oxford Medical Publications, Oxford. 1997.

Briggs AH, Wonderling DE, Mooney C, Pulling cost-effectiveness analysis up by its bootstrap: a non-parametric approach to confidence interval estimation, Health Economics, 1997;6:327-340.

Briggs A and Fenn P, Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane, Health Economics, 1998, 7:723-740.

Briggs A, Gray A, The distribution of health care costs and their statistical analysis foe economic evaluation, Journal of Services Research and Policy 1998; 3(4): 233-245.

Chaudhary MA, Stearns SC, Estimating confidence intervals for cost-effectiveness ratios: an example from a randomised trial, Statistics in Medicine 1996;15: 1447-1458.

Claxton K, Posnett J, An economic approach to clinical trial design and research priority setting, Health Economics 1996;5: 513-524.

Coyle D, Drummond M, Trials and tribulations: conducting economic evaluations alongside trials. Paper presented for the HESG, Winter 1994.

Drummond M, O'Brien B, Stoddard GL, Torrance GW, Methods for the Economic Evaluation of Health Care Programmes (2nd ed) , Oxford Medical Publications, Oxford. 1997.

Greenhalgh T, How to read a paper: Statistics for the non-statisticians. I: Different types of data need different statistical tests, BMJ 1997;315:364-366.

Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. Lancet. 1993;342(Nov27):1317-22

Hair JF, Anderson RE, Tatham RL, Black WC, Multivariate Data Analysis, Prentice Hall, 1984.

Kendall MG and Stuart A, The Advanced Theory of Statistics, Volume I: Distribution Theory. Third edition, London, Charles Griffin & Company Limited, 1969.

Little RJA and Rbin DB, Statistical Analysis with Missing Data. New York, John Wiley and Sons, 1987.

Mainly BFJ, Randomisation, Bootstrap and Monte Carlo Methods in Biology, Chapman & Hall, 1997.

NHS, Improving outcomes in breast cancer – the manual, NHS Executive, 1996.

Obenchain RL, Melfi CA, Croghan TW, Bueshing DP. Bootstrap analyses of cost effectiveness in antidepressant pharmacotherapy. PharmaEconomics, 1997, 11: 464-72.

Rubin DB (1996), Multiple imputation after 18+ years, Journal of American Statistical Association, 91, 413-89.

Sculpher M, Drummond M, Buxton M, The iterative use of economic evaluation as part of the process of health technology assessment, Journal of Health Services Research and Policy 1997:2(1): 26-30.

Severns JL, De Boo T, Knost EM, A comparison of Fieller and bootstrap confidence intervals, International Journal of Technology Assessment in Health Care 1999;15:3. .

Wolstenholme J, Briggs A. Missing….presumed at random: cost-analysis of incomplete data. Paper presented for the HESG, Summer 1999.