An educational piece reviewing the statistical issues in analysing utility data for cost-utility analysis

This is a non-final version of an article published in final form in “Hunter, R. M., Freemantle, N., Baio, G., Butt, T., Morris, S., & Round, J. (2015). An Educational Review of the Statistical Issues in Analysing Utility Data for Cost-Utility Analysis. PharmacoEconomics, 33(4), 355-366. doi:10.1007/s40273-014-0247-6”

Rachael Maree Hunter[1], Gianluca Baio[2], Thomas Butt[3], Stephen Morris[4], Jeff Round[5], Nick Freemantle[1]

1. Research Department of Primary Care and Population Health, University College London (UCL), Royal Free Campus, Rowland Hill Street, London, UK, NW3 2PF.

2. Department of Statistical Science, UCL, 1-19 Torrington Place, London, UK, WC1E 7HB.

3. UCL Institute of Ophthalmology, 11-43 Bath Street, London, UK, EC1V 9EL.

4. Department of Applied Health Research, UCL, 1-19 Torrington Place, London, UK, WC1E 7HB.

5. UCL Clinical Trials Unit, First Floor, 175 Tottenham Court Road, London, UK W1T 7NU, UCL

Corresponding author:

Rachael Maree Hunter

Research Department of Primary Care and Population

University College London Medical School

Royal Free Campus

Rowland Hill Street

London NW3 2PF

Tel. 020 7794 0500 ext 31019

Fax. 020 7670 4890

r.hunter@ucl.ac.uk

Abstract

The aim of cost-utility analysis is to support decision making in healthcare by providing a standardised mechanism for comparing resource use and health outcomes across programmes of work. The focus of this paper is the denominator of the cost-utility analysis, specifically the methodology and statistical challenges associated with calculating quality adjusted life years (QALYs) from patient level data collected as part of a trial. We provide a brief description of the most common questionnaire used to calculate patient level utility scores, the EQ-5D, followed by a discussion of other ways to calculate patient level utility scores alongside a trial including other generic measures of health-related quality of life and condition and population specific questionnaires. Detail is provided on how to calculate the mean QALYs per patient including discounting, adjusting for baseline differences in utility scores and a discussion of the implications of different methods for handling missing data. The methods are demonstrated using data from a trial. As the methods chosen can systematically change the results of the analysis it is important that standardised methods such as patient level analysis are adhered to as best as possible. Regardless researchers need to ensure that they are sufficiently transparent about the methods they use so as to provide the best possible information to aid in healthcare decision making.

Key points for decision makers

- It is best practice for analysts to be open and transparent about how utility scores are used to calculate quality adjusted life years (QALYs) for cost-utility analysis. This includes reporting the questionnaire used, the tariff or method applied to calculate utility scores and for which country, how the issue of missing data has been addressed and if the results have been adjusted for differences in baseline utility scores.
- QALYs should be calculated as area under the curve, discounting for time horizons longer than a year and adjusting for baseline differences using regression analysis. Calculating QALYs using individual level data is more methodologically rigorous than calculating group means at different time points.
- Multiple imputation is considered the most appropriate way to handle missing data.

1. Introduction

The aim of cost-utility analysis is to aid in decision analysis in health care. In particular, it provides a standardised mechanism for comparing resource use and health outcomes across health care technologies and disease areas. The term was coined to differentiate it from cost-effectiveness analysis: a health economic evaluation where the denominator of the incremental cost-effectiveness ratio (ICER) is a cost per unit change in a disease or programme specific outcome, for instance cost per depression free day or cost per infection prevented. Instead, the outcome in the denominator of the cost-utility analysis ICER is a compound measure of mortality and morbidity quantified using preferences or risk in a standardised way [1]. The quantity and quality of life is then combined to calculate quality-adjusted life years (QALYs), disability-adjusted life years (DALYs) or a variant thereof. As the units of analysis are standardised they theoretically make it easier to compare the health and social care resource implications of different programmes and disease areas. Also to note is that cost-utility analysis is still called cost-effectiveness analysis in the US and in the UK a cost-utility analysis may be referred to as a cost-effectiveness analysis or cost-utility analysis depending on the author's preference [1].

In health technology assessments in developed countries QALYs are the most commonly used outcome in the denominator of the cost-utility analysis. QALYs are calculated by weighting each year of life lived using a utility score. Utility scores are anchored so that 1 is perfect health and 0 is equivalent to the state of death. In some models negative scores are possible, representing states that are theoretically worse than death. Multiplying time in a health state by the health state utility value, one year of life lived in perfect health is equal to 1 QALY. If a person were to live for 2 years in a health state that is weighted as 0.5 of full health this is also equivalent to 1 QALY. The utility value and hence QALYs are valued independently of a person's age, so 1 QALY is the same for someone who is 18 years of age as for someone who is 80 years of age [2].

The term "utility" in cost-utility analysis and its theory is based on von Neumann-Morgenstern (vN-M) utility theory. The normative model for utility theory, the model for how a rational individual ought to behave, is that utility scores represent the strength of an individual's preference when faced with uncertainty for a given outcome, in this case a health state. There is a number of conditions that utility scores should meet, and although the model is normative, it should somewhat reflect the way individuals make decisions when faced with uncertainty (some empirical research however suggests these conditions may be more often violated than met [2]). VN-M utility theory also assumes that utility scores are cardinal in that individuals are able to quantify the extent to which they prefer one health state to another. This is as opposed to scores being purely ordinal; individuals are only able to order health states in terms of preference. Theoretically, utility scores that form the basis of QALYs are meant to have these qualities, making them as close as possible to the utility scores in vN-M utility theory [1]. The area though is not without its controversies and the terminology used can be confusing. "Utility" in vN-M utility theory and in the calculation of QALYs should not be confused with the way that the word utility is used in other areas of economics such as welfare theory and Pareto optimisation [2].

Although the original purpose of cost-utility analysis was to aid in decision making related to the allocation of scarce health care resources, the utility scores used to calculate QALYs have found their way into more general analyses of health and social care programmes, with statisticians and economists alike using them to draw conclusions about the effectiveness of

health and social care interventions, policies and technologies. Nevertheless, there is a range of methods that analysts can use to calculate QALYs and account for missing data. A literature review by Richardson and Manca [3] identified a number of different ways way that QALYs were calculated in cost-utility analyses alongside randomised control trials. They identified that the calculation of QALYs in many cases lacked consistency and that there was poor reporting of methods. Reporting of missing data and how it was handled were of particular concern, with almost a half of papers not reporting the amount of missing data and only one fifth with missing data including and describing a method of data imputation. The authors recommend that cost-utility analyses need to be fully transparent with consistent analytical techniques employed, as different methods can influence the direction and significance of the results [3]. As utility scores gain increasing appeal with a wider range of analytical disciplines the risk of miscommunication and misinterpretation of analyses of utility scores increases.

The focus of this article is the denominator of the cost-utility analysis, specifically the calculation of QALYs in the incremental cost per QALY gained analysis. We aim to re-orientate utility scores towards their original standardised purpose and provide guidance to analysts including statisticians, epidemiologists and junior health economists, new to economic evaluation on how to calculate QALYs from utility scores collected as part of a trial, highlighting specific statistical issues that are important to address. We firstly provide a brief description of the most common questionnaire used to calculate patient level utility scores, the EQ-5D, followed by a discussion of other ways to calculate patient level utility scores alongside a trial including other generic measures of health-related quality of life and condition and population specific questionnaires. In section five we go into detail on how to calculate the mean QALYs per patient using patient data that has been collected as part of a trial. Building on the formulas set out in section five we describe the methods for discounting and adjusting for baseline differences. In the final two sections we describe different ways to handle missing data including multiple imputation in section eight.

2. Utilities from the EQ-5D

Calculating patient level utility scores as part of a trial is usually made up of two steps: (1) asking patients to complete a generic measure of health-related quality of life at different time points over the duration of the trial, including baseline to measure a patient's health status; and (2) applying a preference based algorithm to calculate utility scores for each patient's health status at each time point.

The key body responsible for providing advice and guidance on best clinical practice including value for money in England (the National Institute for Health and Care Excellence, NICE) recommends that utility scores are obtained from a random sample of the general population and using a technique called time trade-off (TTO) to arrive at a measure of preference under uncertainty for a given health state [4]. The generic questionnaire most favoured in the UK is the EuroQol group's EQ-5D and the variant currently most commonly used is the 3 level version. This questionnaire consists of five questions asking if patients have no, some or extreme problems with mobility, self-care, usual activities, pain and anxiety/depression. In total the questionnaire defines 243 distinct health states, each of which have an associated, country specific, utility score representing the preferences of a sample of the general population of that country. In the UK the utility scores for the EQ-5D 3 level ranges from 1 for perfect health to -0.594 - states worse than death are possible with

this value set, which are given negative scores [6]. A 5 level version of the EQ-5D is available and preference based health state valuations from a random sample of the general population are in the process of being derived [7].

It should be noted that the issues discussed in this paper are not specific to utility scores derived from the EQ-5D but also to other generic measures of health-related of quality life and condition and population specific measures that have preference based algorithms and can be used to calculate utility scores for health states. The other methods will be discussed in the next two sections. If the method used is something other than using a country specific version of the EQ-5D to calculate QALYs, caution should be exercised when comparing the results of this analysis with others as the results will systematically vary from method to method [8].

3. Calculating utility scores using other generic measures

In some instances the view is taken that the EQ-5D is not suitable for a particular patient group or disease [9]. It is also the case that in some trials EQ-5D data is not collected. Data from other generic measures of health-related quality of life might have been collected though. Brazier et al. [10] developed a 6 dimension classification of another measure of generic health-related quality of life that is commonly used in trials: the short form 36 and short form 12 (SF-36 and SF-12). The 6 dimension classification system, commonly referred to as the SF-6D, then has utility values from members of the general public using the standard gamble and an algorithm to convert either the SF-36 or SF-12 into a single utility score. Another example includes the multi attribute health status classification system: Health Utilities Index (HUI). The HUI has associated utility scores derived from a Canadian population [11]. The HUI2, which is designed to be used in children, has a UK algorithm for calculating utility scores [12]. Another generic health-related quality of life questionnaire for use in children with an associated UK preference set is the Child Health Utility 9D (CHU 9D) [13].

4. Calculating utility scores using condition specific measures and mapping.

In addition, criticism has been levelled at some of the measures as they do not capture domains of interest, for example there are no questions in the EQ-5D directly relating to vision, fatigue or cognition [9]. To attempt to address this issue new utility preference sets based on TTO or SG have been generated from a condition specific questionnaire. This includes in cancer, where a specific utility value set has been generated for the commonly used EORTC-QLQ-C30 [14] and in epilepsy where a utility value set has been developed for the NEWQOL-6D questionnaire for epilepsy [15]. Questionnaires and utility value sets have also been generated for specific populations that might struggle to fill in the traditional questionnaires or where extra domains where considered suitable. For example, to address the challenges of QoL measurement for patients with dementia the DEMQoL [16] and its associated value sets [17] have been developed. It is recommended that these questionnaires are used alongside the EQ-5D for consistency, and as it is currently unclear what impact that might have on utility estimates [17]. Another issue, particularly pertinent to patients where cognitive ability is limited or they are too unwell to complete the questionnaire, is completion of questionnaires by proxies compared to self-completion. The EQ-5D has a proxy completed version, although the validation of this has been limited and the viewpoint of the proxy is an important determinant of how the questionnaire is completed

and the resulting utility scores [18]. Researchers concerned about consistency are advised to use both proxy and patient completion where possible [17]. This is not always possible though and in these instances the person completing the questionnaire should be noted (individual or proxy) and a sensitivity analysis conducted.

Another solution for calculating utility scores when EQ-5D data has not been collected is to use a mapping algorithm derived from "mapping" a generic preference based measure of health-related quality of life onto a condition specific questionnaire. Commonly used condition specific measures with a mapping algorithm are the European Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire EORTC-30 in cancer [19] and the Barthel index in stroke care [20]. Extreme caution needs to be exercised when using mapping algorithms though as in many cases they can over or under predict significantly for some patients, partially because of the methods used to derive the algorithms [21].

A more thorough discussion of the issues of using a condition specific instrument for calculating utility scores can be found in Brazier et al [22].

5. Calculating a QALY

The following sections use trial data to illustrate how to calculate QALYs and the impact that different methods and assumptions can have on the direction and magnitude of the results of an analysis. Following the collection of patient health status as part of the trial, and once the algorithm for that measure is applied to calculate utility scores for patients, the most commonly accepted way to calculate QALYs is to calculate the area under the curve (AUC) joining the utility measurements at each point in time [1]. To illustrate the calculation of QALYs we use trial data from real patients with data on follow-up time points of 6 months, 1 year and 2 years for 152 patients (85 in the treatment group and 67 in the control group) [23][1]. The treatment arm involved a health check at baseline for patients with intellectual disabilities; the control group received usual care. Figure 1 and columns 2 and 3 of table 1 show mean utility scores calculated from the EQ-5D over the two year period for the treatment and control groups and the numbers with complete EQ-5D value sets in table 2. The amount of missing data for the EQ-5D at each follow-up point is reported in table 2.

There are two broad ways that the AUC can be calculated: (i) at the patient level using patient level data at each follow-up point; (ii) at the group level using the mean utility score for each follow-up point.

The recommended method is to calculate QALYs using individuals' utility scores at different time points.

$$q_{jti} = \frac{(u_{j(t-1)i} + u_{jti})}{2}\delta_t \qquad (1)$$

[1] Original follow-up points for trial data were 3, 6 and 9 months. These have been extended so as to demonstrate the impact of discounting which only occurs after 1 year.

The formula to calculate the AUC or QALYs using patient level data is shown above in Formula 1, where *u* is the utility score, *i* denotes an individual, and *j* is time so that at baseline *t*=0. For each group *j* (*j*=0 for control and *j=1* for treatment) the consecutive time measures are added, averaged and then re-scaled (δ) for the percentage of a year that *t* and *t* - 1 cover, so 0.5 for 6 months or 0.25 for 3 months and so on.

$$Q_{ji} = \sum_{t=1}^{T} q_{jti} \tag{2}$$

$$Q_j = \frac{\sum_{i=1}^{n} Q_{ji}}{n} \tag{3}$$

For the total duration of the trial, the total QALYs (Q) for each individual are the summation of the QALY calculations for each follow up time point starting at *t*=1, the first follow up point (formula 2). The mean QALYs for each treatment group (Q_j) are then calculated from the individual level data, dividing the sum of all QALYs for all patients ($\sum Q_{ji}$) by the number of patients, (n) (formula 3). Standard deviations, standard errors and confidence intervals can then be calculated using standard formulas from the individual level data in formula 2. Most statistical packages will omit individuals with any missing data on the questionnaire or utility score at any time point from the analysis if no adjustment to the analysis is made to account for this. Hence the resulting analysis will be an individual level, complete case analysis (only individuals who have utility scores at baseline and all follow up time points will be included). In the example data set at the final follow-up point, 24 months, 85% of participants have complete responses to the EQ-5D in the control group and 83% in the treatment group. Across all time points though, only 70% of participants in the control group and 71% of participants in the treatment group have complete responses to the EQ-5D (see table 2). When an individual level, complete case analysis is run using the example data set the total mean QALYs for the control group are 1.47 (95% CI 1.36-1.57) and for the treatment group are 1.39 (95% CI 1.27-1.51) (see table 3).

When reporting the results of the analysis the mean utility score and standard deviations for each follow-up point might be reported to provide more detailed information. If this is done then care needs to be taken in reporting and interpreting these results. The mean utility for each treatment group (u_{jt}), using all individuals with complete data at each time point, an available case analysis, was calculated using the example data set and is reported in Table 1. If these available case means are used to calculate the area under the curve adjusting for timing (δ) (formula 4) then summed for the total duration of the trial (formula 5) the results differ from the individual level analysis as additional individuals are included in the analysis. The total QALYs are 1.37 for the control group and 1.35 for the treatment group. A complete breakdown of the mean utility scores at each follow-up point and total mean QALYs as an available case analysis is reported in table 1 for illustration purposes.

$$q_{jt} = \frac{(\bar{u}_{j(t-1)} + \bar{u}_{jt})}{2} \delta_t$$

(4)

$$Q_j = \sum_{t=1}^{T} q_{jt}$$

(5)

Note that the results are different for the individual level complete case analysis compared to the available case analysis. If a complete case analysis is conducted using mean follow-up scores at the group level, so only including individual with complete questionnaires at all points, the results are the same as at the individual level (see table 3). It is important to note that both of these examples are for illustrative purposes only, so as to provide a description of the calculation of QALYs. Neither complete case analysis nor available case analysis adequately addresses the issue of missing data and hence both of them are biased. This will be discussed further in section 8 below.

Both of these analyses also assume that the line drawn between utility scores and hence QALYs are linear. This significantly simplifies the analysis as the assumption of linearity allows two consecutive utility scores to be averaged, and then multiplied by the duration between the two scores. Some authors though have experimented with non-linear changes between consecutive time points (e.g. Miyatmoto [24]). These models are considerably more complex and, although they will report different results to a linear model, will only make a minimal difference to the results over short trial durations or if follow-up time points are close. Authors such as Billingham et al [25] have also investigated assumptions regarding changes over time and mortality using survival analysis provided recommendations for calculating QALYs.

The SF-36 was also administered to patients in the example trial at baseline and the last follow up point, 2 years. Applying the SF-6D algorithm developed by Brazier et al [10] and doing a complete case analysis the mean QALYs for the control group are 1.51 (95% CI 1.45 to 1.57) and for the treatment group 1.54 (95% CI 1.48 to 1.60).

Note that the final score for the SF-36 and associated SF-6D algorithm is higher than that of the EQ-5D. This is common for studies using the SF-6D, partially because the range of utility scores for the two algorithms are not the same. Whereas the range of possibility utility values for the EQ-5D ranges from 1 to -0.594, with a number of health states that have utility scores worse than death, scores on the SF-6D are anchored at 1 for perfect health but with a minimum score of 0.316. As a result the SF-6D tends to be worse at discriminating between responses for patients in worse states of health. At the upper ranges though it tends to discriminate better, having less of a ceiling effect than the EQ-5D [26]. If using the SF-6D rather than the EQ-5D for an analysis one must be careful of this systematic bias when interpreting the results of the incremental cost effectiveness ratio.

6. Discounting

In a cost-utility analysis both costs and QALYs need to be adjusted for the different times that they occur. Discounting is the process of accounting for time preferences in the receipt of costs or benefits. The standard discount rate used by NICE for both costs and benefits is 3.5% [4], with a standard of 3% being consistent across other developed countries including the US [27]. For public health interventions NICE recommends a discount rate for costs and benefits of 1.5% [28]. Discounting can be done using group means or using individual level data, although it is recommended that discounting is done at the individual level, particularly due to issues associated with missing data, as will be discussed in section 8.

$$u_{jti} = \frac{u_{jti}}{(1+r)^{(t-1)}} \tag{6}$$

Formula 6 is the discount formula for calculating QALYs at an individual level using the utility score (*u)* for individual *i* at time *t* for a discount rate of *r*. The discount rate is generally only applicable in cases where the duration of the trial is greater than 12 months. In this formula *t* is the time from baseline in years. It is possible though to do a monthly discount rate by converting the yearly discount rate *r* to a monthly value (Formula 7). The same discount rate formula still applies, but instead using the new 1 month rate and *t* becomes the time from baseline in months. In the case of a discount rate of 3.5%, 0.3% would be the monthly discount rate using Formula 7 below.

$$r_{month} = (1+r)^{\frac{1}{12}-1} \tag{7}$$

Discounting has been applied to the example data set at the group level (available case analysis) to provide a worked example (see table 1). When applied at the individual level (complete case analysis) the new mean QALYs and 95% CI are 1.44 (95% CI 1.34-1.54) and 1.37 (95% CI 1.25-1.48).

7. Adjusting for baseline differences

In trials random allocation to the treatment or control group is the most common mechanism used to try to ensure balanced characteristics between treatment and control groups. Randomisation though is not a perfect mechanism for this and sometimes differences can occur between groups for patient characteristics that might have a significant impact on the results of the trial. For example, by chance, individuals in the control group may have more severe disease symptoms at baseline than individuals in the treatment group. Hence, other additional methods are sometimes used to try to minimise the impact that differences between groups can have on the results of the analysis, some of which can occur prior to randomisation, such as stratification. Stratification is where patients are randomised to intervention groups in such a way as to try to reduce significant differences between groups for a specific important characteristic, such as co-morbidities or symptom severity. Stratification can also be used in very specific situations, such as controlling for the random allocation to treatment groups between countries in a multi-country trial.

Another common technique, particularly in observational data, is adjusting for baseline characteristics where there is strong reason to believe that they could influence the outcome of interest. Adjusting for baseline differences in groups does come with caveats. Generally including variables used for stratification as covariates in a regression analysis is considered acceptable, and in some instances baseline variables that are known *a prior* to be strongly or moderately associated with the dependent variable in the analysis. However, it is important to note that variables that were measured after randomisation or that could be influenced *post hoc* by the treatment in any way should not be included as covariates for the purposes of baseline adjustment and the variables chosen should also be specified in the analysis plan [29].

In the calculation of QALYs imbalances in baseline values has a different implication for the analysis. QoL is not a variable that is included in stratification and is generally a dependent variable, not a covariate in an analysis, so it is not intuitively adjusted for. It is possible though that the treatment or control group may differ in QoL or utility scores at baseline. The nature of calculating QALYs as AUC then means that one group may have more QALYs by virtue of their baseline utility scores alone, particularly if the treatment only has a minimal effect on QoL. In the example data set, by chance, the utility score for the control group is higher than for the treatment group, 0.68 (95% CI 0.61-0.74) versus 0.63 (0.57-0.70) respectively, although not significantly higher.

There are a number of methods that can be used to adjust for baseline utility score differences, with linear regression with baseline adjustment being the recommended method [29].

One approach that has commonly been used, and is the most intuitive, is to calculate the difference in utility score from baseline, so that only the actual change in utilities is computed and baseline differences are removed. Looking at the group means in table 1 only, in the control group there is an increase of 0.04 from baseline over the 2 years and for the treatment group an increase of 0.17. Using the individual level data in the example data set, the mean difference in utility score from baseline for the control group is -0.02 (95% CI -0.8 to 0.4) and for the treatment group 0.16 (95% CI 0.09 to 0.23), with the increase in utility score being significantly different in the treatment compared to control group. This methodology though presents significant problems for calculating QALYs for hypothesis testing as only the less informative mean difference from baseline is calculated. This method has also been criticised in the literature as it fails to control for imbalances due to regression to the mean; patients' future scores are generally correlated with their past scores [30].

Instead the method that has been recommended is to use linear regression with baseline adjustment. In this instance QALYs for each individual are calculated as described in section 5 above. A linear regression analysis is then run where the dependent variable is the total QALYs, β_1 is the coefficient for the treatment effect and β_2 the coefficient for baseline utility scores:

$$Q_i = \beta_0 + \beta_1(j_i) + \beta_2(u_{i,t=0}) \quad (8)$$

In this linear regression analysis the coefficient for β_1 is then the difference in QALYs between treatment and control. In a standard statistics package like Stata this can be done

by running the regression analysis (`regress total QALYs treatment_dummy_variable baseline_utility`). Running this analysis on the example dataset the treatment co-efficient, β_1, is 0.029 (95% CI -0.067 to 1.262). The mean, adjusted QALYs for each treatment group can also be calculated using most standard statistical packages. In Stata this is done by using the command `adjust` after running the regression analysis, and adjusting for baseline utility scores at mean (`margins treatment_group, at(baseline_ utility=mean)`). If this is done in the example dataset the mean patient level QALYs for the control group are 1.35 (95% CI 1.28 to 1.43) and for the treatment group they are 1.38 (95% CI 1.32 to 1.45) undiscounted. Including discounting they are 1.33 (95% CI 1.26 to 1.40) and 1.36 (95% CI 1.3 to 1.42) (see table 3). Note that this method only adjusts for baseline utility scores and not other baseline characteristics, which as mentioned above, is to be done with caution unless they are stratified variables or they are pre specified in the analysis plan.

In this example, although adjusting for baseline differences does not change the conclusion (there is a non-significant difference between treatment and controls even after adjusting for baseline differences) it does change the direction of the difference from negative to positive. Although this does not change the conclusion of the analysis, there is no difference in QoL between the two groups, it may have an impact on the results of a cost-utility analysis once costs are added in. This is discussed further in section 9.

Other statistical methods of varying levels of sophistication, particularly those that are designed to account for a range of baseline patient characteristics, are also possible. An example would be to fit a multi-level model with fixed and random effects. One option, which is explained more fully in Carpenter and Kenward [31], is to account for the repeated measures of utility within a patient as the response variable and include a random effect accounting for baseline response at the patient level. This approach has the statistical advantage of parameterising the baseline measurement (eg accounting for measurement error on that value) and enabling analysis of data that are incomplete (eg subjects with say, only endpoint values, may be included). Fitting this multi-level model assuming a fixed effect for the treatment coefficient, the estimated mean undiscounted QALYs for the control group is 1.55 (95% CI 1.47 to 1.62) with an estimated reduction of QALYs for the treatment group of -0.0029 (95% CI -0.11 to 0.11).

8. Missing data

Missing data is a well-documented problem in health care data sets and methods to address it will almost always need to be considered within the statistical and economic evaluation analysis plan for any repeated measures patient level data set. In the first instance though it can be considered a trial design issue, where by trialists and research staff should ensure that the methods and resources they use to recruit and retain patients try to minimise the risk of missing data as best possible. In terms of costs for economic evaluations this can be addressed by taking data from medical records. For questionnaires that are not routine though this presents more of a challenge. Length of the questionnaire has been shown to have a negative correlation with probability of responding so shorter and fewer questionnaires are preferrable [32]. In this way the EQ-5D can be a good questionnaire given its relatively short length.

Once the data have been obtained, and the level of missingness ascertained, it is up to the analysts to handle the missing data appropriately. This would include an assessment of the mechanism of missingness: Is the data missing completely at random (MCAR)? This is where there is no relationship between the data that is missing and any values in the data set, missing or observed. Is the data missing at random (MAR)? This is where the missingness is not related to the missing data but an observed variable in the dataset. Or the missing data can be missing not at random (MNAR) where the propensity for data to be missing is related to the missing data. When data are MAR observed data can be used to estimate the results. There are different ways to test for which observed variable predicts if data are MAR, which are covered more fully in Little and Rubin [33], but once the mechanism for missingness is ascertained the most appropriate method for handling missing data needs to be chosen, taking into account the data available. It should be noted though that there is no way to test if data are MCAR given that there is chance that the propensity for data to be missing is as a result of some individual characteristic that is missing from the data set.

A recent systematic review of economic evaluations along randomised control trials (RCTs) found that missing data was rarely handled appropriately and in a fifth of cases it was unclear if it had been accounted for or was not mentioned at all. In a third of cases complete case analysis was the analysis method chosen [34]. Complete case analysis though is considered a poor statistical technique for dealing with missing data as it significantly biases the results. It makes the assumption that individuals whose data is missing within a data set have the same characteristics as those with complete data; that the data is missing completely at random. In most clinical trials this is unlikely to be the case. The most common bias is that individuals whose disease is more severe or who do not respond as well to treatment are more likely to have missing data, as they may be more unwell or have lost interest in the trial given it has provided them with no benefit. Individuals who fail to comply with treatment are also less likely to show up for further assessments and hence are more likely to have missing data.

The level of missingness in the example data set at each follow up time point is reported in table 2. At 24 months 57 individuals have complete data in the control group and 71 in the treatment group. This represents approximately 85% of the participants in both groups. The main analysis used thus far has been a complete case analysis. Only 47 individuals in the control group and 60 in the treatment have complete data over the full duration of the trial, representing 70% of participants. To determine if there is an observable mechanism by which data are missing comparisons can be made between fully observed baseline characteristics, such as age, gender and intellectual ability, and the missing outcome, utility scores. One can then define a variable R, where R=1 if utility scores are observed at all follow-up points for an individual, or R=0 if there are any missing observations. It is then possible to test for differences in the rate of missingness (R) for fully observed baseline variables using simple statistical tests such as Chi-Square for categorical observed variables and T-tests for continuous. Although no relationship in the example data set was found between level of learning disability (mild, moderate, severe or profound) and missingness or gender and missingness, age was significantly related to the amount of missing data, with the older the individual the more likely they had missing data. EQ-5D utility scores at baseline were also significantly lower for individuals with missing data. In this instance one

can hypothesise that data is not missing at random, but that older individuals with lower quality of life are less likely to return subsequent questionnaires.

There are a number of ways that one can work with missing data, although only three will be dealt with here; (i) taking the last utility value forward; (ii) taking the last questionnaire value forward; and (iii) multiple imputation (MI). These particular methods have been chosen because they illustrate two types of missing data analysis (simple and random) and the impact different methods can have on the outcome of the analysis. Taking the last value forward, examples i and ii, can be considered a simple approach to missing data. Another simple approach to missing data, that will not be addressed here, is imputing mean values. This particular option has not been chosen as from the previous analysis above it is clear that individuals with missing values vary systematically to those who do not, specifically having lower baseline utility values. Imputing mean results for follow up scores would bias the results of the analysis as it would potentially artificially inflate the follow-up utility scores. There are more ways to deal with missing data than the examples given. A more complete review of the literature is available in Eekhout et al [35]. As a general rule though MI is the most appropriate technique to use when analysing patient level datasets with missing data. A more complete tutorial on MI methods is also available in White et al [36].

8.1 Last value forward: utility score

An assumption one might make when calculating the AUC is that the quality of life (QoL) or utility score of patients with missing data at point *t* is the same utility score they experienced in the previous time point (*t*-1). This essentially draws a straight, horizontal line from the last point with available data to the next point where data is available to calculate the area under the curve. It does introduce bias though in that if the reason individuals that were unable to complete questionnaires are because their health deteriorated and were too unwell to reply the values for these individuals will be artificially inflated.

If this assumption is applied to the example data set then the mean QALYs for each group using a patient level analysis (formula 5)) the mean unadjusted, undiscounted QALYs are 1.35 (95% CI 1.22 to 1.46) for the control group and 1.35 (95% CI 1.25 to 1.46) for the treatment group; including discounting and adjusting for baseline values (formula (8)) the mean QALYs are 1.3 (95% CI 1.22 to1.37) for the control group and 1.35 (95% CI 1.28 to1.42) for the treatment group.

8.2 Last value forward: questionnaire response

In some instances utility scores cannot be calculated because not all items of the questionnaire are completed, rather than the whole questionnaire being missing; this is sometimes referred to “item-non response” (as opposed to unit- non response, where the whole questionnaire is missing). In the example data set, one item missing from the questionnaire accounted for a significant proportion of the missing utility scores at 6 and 24 month follow ups. Similar to the previous method, the last value given for that question in the previous time-point can be assumed to be the same in this instance and EQ-5D scores calculated from there.

If this assumption is applied to the example data set then the mean QALYs for each group using a patient level analysis (formula 5)) the mean unadjusted, undiscounted QALYs are 1.34 (95% CI 1.23 to 1.46) for the control group and 1.35 (95% CI 1.25 to 1.46) for the treatment group; including discounting and adjusting for baseline values (formula (8)) the

mean QALYs are 1.3 (95% CI 1.22 to 1.37) for the control group and 1.35 (95% CI 1.28 to 1.42) for the treatment group.

The values are very similar to those where utility is replaced. As a result one can argue that it doesn't matter whether you replace missing utility score or questionnaire values. On the other hand, the main underlying assumption behind this strategy is that given the past history, missingness is completely non informative. Thus, it is reasonable to assume that patients will continue to show the same behaviour as they have in the past. In most cases, this seems too stringent an assumption, thus potentially limiting the usefulness of these methods. In this instance given that the patients with the lowest utility are those that are the most likely to have missing values, using either of the last value forward methodologies does nothing to address the systematic bias associated with the missingness of the data. In any case, it should be pointed out that the results are strongly dependent on this assumptions and thus extensive sensitivity analysis is recommended.

8.3 Multiple Imputation

One method that usually overcomes the severe limitations of last value forward and other simple imputation methods is that of stochastic multiple imputation. The basic idea is to "impute" the values that have not been observed with some summary obtained using the available data. Typically, this is done by taking random draws from a suitable probability distribution to account for the underlying uncertainty in the summary statistic. This random draw is used to represent the potential observation that just happens not to be available. A distinct characteristic of analyses based on multiple imputation is that rather than having a singular dataset where missing values are replaced by a new value, as is the case with simple imputation methods, for multiple imputation the original dataset remains as is, along with the missing values, and a number of new datasets are created where a new value is inserted in place of the missing one. In each new dataset the new imputed value varies between datasets, within likely ranges for the parameter and can be based on pre specified relationships with observed baseline characteristics, as will be described below. In summary, a number of new hypothetical datasets are created with different values imputed for the missing data, leaving the original dataset intact. The same analyses as previous are then run on the new datasets, which are combined using appropriate methods, to obtain an estimate of the parameter of interest, the standard error and the standard deviation.

The selected probability model describes the joint variation of observed and unobserved values. One simple way of doing so is to set up a regression model (on a suitable scale), in which the variable for which missing data are present depends on some fully observed covariates. A by-product of this procedure is that it automatically takes into account the patient characteristics that are a) available to the researcher; and b) considered to be predictive of the missing mechanism. Of course, from the computational point of view, this implies that the analysis is generally more complex and requires more advanced statistical tools.

To choose the number of imputed datasets to create, a general rule of thumb is to create *m* imputed datasets such that *m* is equal to the number of percentage points of missing data for the dependent variable [36]. In the example data set, across all time points a utility score was missing for 30% of data points, and hence 30 imputed datasets were created. Previously, when less computing power was available to run analyses it was generally recommended that three to ten datasets were created, with limited extra benefit obtained

after ten [33]. Given the additional computing power available today to run analyses even generating hundreds of imputed datasets will have little impact on the time taken to run analyses on relatively small trial datasets. To test though that an adequate number of datasets have been generated White et al [36] suggested the following additional tests:

1) That the Monte Carlo error of the regression coefficient for the treatment variable, β_1 in formula 8, is approximately 10% of its standard error.
2) The Monte Carlo error of the test statistic for β_1, is approximately 0.1.

They set a third assertion for the P-value given that the assumption is that p should be less than 0.05. Given that trials are rarely powered to find a difference in QALYs between treatment and control groups it is unlikely that there will be a difference between the two groups and hence this condition is rarely relevant for QALYs. The Monte Carlo estimate can be calculated in Stata using the command `mcerr` following the command `mi estimate`.

The only observed complete baseline variable available that was informative of missingness in the example dataset, as described above, was age. In Stata, while the data was still in long format (multiple observations for each individual in the one column) the data was declared as long (`mi set mlong`), EQ-5D utility scores were declared as the variable to be imputed (`mi register impute utility score variable`), and 30 imputed datasets were created using regression analysis and age as an informative observed variable (`mi impute regress utility score age, add(30)`). The data was then reshaped into wide, using Rubin's rule to combine information on each individual, using the "`mi reshape`" command in Stata and further analyses carried out using "`mi estimate`". A regression analysis run on the imputed data estimated that the mean undiscounted and unadjusted QALYs for the control group were1.36 (95% CI 1.25 to 1.47) with the treatment group having 0.02 QALYs less on average (95% CI -.17 to 0.127). Adjusting for baseline differences in utility scores the treatment group had a QALY increase of 0.02 (95% CI -0.089 to 0.129).

9. Probability distribution

The examples above do not address the distribution of utility data. Utility scores derived from the EQ-5D have a number of distribution issues in that, in addition to having the qualities of a maximum value of 1 and minimum value of -0.594, EQ-5D utility scores tend to be positively skewed with an identifiable ceiling effect[2]. Figure 2 is the frequency distribution of the EQ-5D utility data at baseline from the example dataset. It is clear from the graph that EQ-5D utility scores in this example are not normally distributed. The most common way of addressing this in cost-utility analysis is to use non-parametric tests of significance, with the recommended method being bootstrapping [27] where N samples of n individuals are taken from the original dataset, and then replaced. The process is then repeated a large number of times (upwards of 1,000), and for each of these samples the mean of the sample is calculated. Further statistics can then be run on these results. Additionally, if these means are then stored in memory, this has the added advantage of alongside bootstrapped cost data cost-effectiveness planes and cost-effectiveness acceptability curves can be created that allow for reporting the percentage of times that the mean of samples for a given option are cost-effective for a willingness to pay for a QALY. Other methods that can take into

[2] Note that some issues with the ceiling effect seen in the EQ-5D have been overcome by the development of a 5 level version of the questionnaire [37]

account the distribution of the data and some of the other issues discussed above are more complex regression techniques and seemingly unrelated equations which can also be used for hypothesis testing of cost and utility data but are beyond the scope of this article [36].

10. Conclusion

Utility scores and the calculation of QALYs present a number of statistical analysis challenges, some of which are due to the nature of clinical trials, such as missing data. This paper has demonstrated that the methods used are important as they can impact on the direction or magnitude of the results, although in this instance not necessarily the conclusions. Use of standardised methods is recommended so as to maintain the original intended purpose of utilities and QALYs – comparison across disease areas and different programmes of work. If this is not possible though analysts need to ensure that they are transparent about their methods and any assumptions made regarding the above areas so that decision makers and others are able to make informed decisions about the effectiveness of a new technology and whether to implement or to conduct further research.

Acknowledgements:

We would like to thank Professor Sally-Anne Cooper and colleagues at the University of Glasgow for allowing us to use their data as part of this analysis. The data was amended for demonstration purposes and so that no comparison can be drawn between the results of this analysis and any results of the trial.

Conflict of Interest

No funding was received for the analysis or writing of this paper. None of the authors have any conflicts of interest to report.

Individual author contributions

All authors contributed to the original idea for the paper and have provided written contributions to the paper including edits to draft versions. RMH wrote the original draft, conducted data analysis and co-ordinated additional edits to the paper. GB contributed to the section on missing data in addition to comments and edits on the paper. TB contributed to sections on alternatives to the EQ-5D in addition to comments and edits on the paper. NF, SM and JR provided expertise on the overall ideas and content of the paper in addition to edits and comments on each draft of the paper.

RMH will act as overall guarantor for the work.

References

1. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. 3rd Edition. Oxford: Oxford University Press. 2005.
2. Torrance GW, Feeny D. Utilities and quality adjusted life years. Int J Technol Assess. 1989;5:559-575
3. Richardson G, Manca A. Calculation of quality adjusted life years in the published literature: a review of methodology and transparency. Health Econ 2004;13:1203–10.
4. National Institute for Health and Care Excellence (NICE). Guide to the methods of technology appraisal 2013. NICE. 2013. http://publications.nice.org.uk/pmg9. Accessed 24 March 2014
5. Oemar M, Oppe M. EQ-5D-3L User Guide. Basic information on how to use the EQ-5D-3L instrument. Version 5.0. EuroQol. 2013. http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/EQ-5D-3L_UserGuide_2013_v5.0_October_2013.pdf. Accessed 24 March 2014.
6. Dolan P. Modelling valuations for EuroQol health states. Med Care 1997;35:1095-1108.
7. Oemar M, Janssen B. EQ-5D-5L User Guide. Basic information on how to use the EQ-5D-5L instrument. Version 2.0. EuroQol. 2013. http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/UserGuide_EQ-5D-5L_v2.0_October_2013.pdf. Accessed 24 March 2014.
8. Brazier JE, Rowen D, Mavranezouli I, Tsuchiya A, Young T, Yang Y, Barkham M, Ibbotson R. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technol Assess 2012;16:1-114.
9. Longworth L, Yang Y, Young T, Mulhern B, Hernández Alava M, Mukuria C, Rowen D, Tosh J, Tsuchiya A, Evans P, Devianee Keetharuth A, Brazier J. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. Health Technol Assess 2014;18:1-224.
10. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271-92.
11. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-Attribute Health Status Classification Systems: Health Utilities Index. Pharmacoeconomics 1995;7:490-502.
12. Stevens K, McCabe CJ, Brazier JE, Roberts J. Multi-attribute utility function or statistical inference models: a comparison of health state valuation models using the HUI2 health state classification system. J Health Econ 2007;26:992-1002.
13. Stevens K. Valuation of the Child Health Utility 9D Index. Pharmacoeconomics 2012; 30:729-747
14. Rowen D, Brazier J, Young T, Gaugris S, Craig BM, King MT, Velikova G. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. Value Health 2011;14:721-31.
15. Mulhern B, Rowen D, Jacoby A, Marson T, Snape D, Hughes D, Latimer N, Baker GA, Brazier JE. The development of a QALY measure for epilepsy: NEWQOL-6D. Epilepsy Behav 2012;24:36-43.

16. Smith SC, Lamping DL, Banerjee S, Harwood R, Foley B, Smith P, Cook JC, Murray J, Prince M, Levin E, Mann A, Knapp M. Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. Health Technol Assess 2005;9:1-93.
17. Mulhern B, Rowen D, Brazier J, Smith S, Romeo R, Tait R, Watchurst C, Chua K-C, Loftus V, Young T, Lamping D, Knapp M, Howard R, Banerjee S. Development of DEMQOL-U and DEMQOL-PROXY-U: generation of preference-based indices from DEMQOL and DEMQOL-PROXY for use in economic evaluation. Health Technol Assess 2013;17:1-140.
18. Bryan S, Hardyman W, Bentham P, Buckley A, Laight A. Proxy completion of EQ-5D in patients with dementia. Qual Life Res 2005;14:107-18.
19. Kontodimopoulos N1, Aletras VH, Paliouras D, Niakas D. Mapping the cancer-specific EORTC QLQ-C30 to the preference-based EQ-5D, SF-6D, and 15D instruments. Value Health 2009;12:1151-7.
20. Kaambwa B, Billingham L, Bryan S. Mapping utility scores from the Barthel index. Eur J Health Econ 2013;14:231–41.
21. Longworth L, Rowen D. Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. Value Health 2013;16:202-10.
22. Brazier JE, Rowen D, Mavranezouli I, Tsuchiya A, Young T, Yang Y, Barkham M, Ibbotson R. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technol Assess 2012;16:1-132.
23. Cooper SA, Morrison J, Allan LM, McConnachie A, Greenlaw N, Melville CA, Baltzer MC, McArthur LA, Lammie C, Martin G, Grieve EAD, Fenwick E. Practice nurse health checks for adults with intellectual disabilities: a cluster-design, randomised controlled trial. Lancet Psych 2014;7:511-521.
24. Miyamoto JM. Quality-adjusted life years (QALY) utility models under expected utility and rank-dependent utility assumptions. J Math Psychol 1999;43:201-37.
25. Billingham LJ, Abrams KR, Jones DR. Methods for the analysis of quality-of-life and survival data in health technology assessment. Health Technol Assess 1999;3:1-152.
26. Whitehurst DG, Bryan S, Lewis M. Systematic review and empirical comparison of contemporaneous EQ-5D and SF-6D group mean scores. Med Decis Making 2011;31:E34-44.
27. Glick HA, Doshe JA, Sonnad SS, Polsky D. Economic Evaluation in Clinical trials. Second Edition. Oxford: Oxford University Press. 2007.
28. NICE. Incorporating health economics. In: Methods for the development of NICE public health guidance (third edition). NICE. 2012. http://publications.nice.org.uk/methods-for-the-development-of-nice-public-health-guidance-third-edition-pmg4/incorporating-health-economics. Accessed 24 March 2014
29. Committee for Medicinal Products for Human Use (CHMP). Guideline on adjustment for baseline covariates. European Medicines Agency. 2013. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/06/WC500144946.pdf. Accessed 17 June 2014
30. Manca A, Hawkins N, Sculpher MJ. Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. Health Econ 2005;14:487-96.

31. Carpenter JR, Kenward MG. Missing data in randomised controlled trials— a practical guide. Birmingham: National Institute for Health Research. 2008. www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml. Accessed 24 March 2014
32. Brueton VC, Tierney JF, Stenning S, Meredith S, Harding S, Nazareth I, Rait G. Strategies to improve retention in randomised trials: a Cochrane systematic review and meta-analysis. BMJ Open 2013;4:e003821.
33. Little RJ, Rubin DB. Statistical analysis with missing data. 2nd Edition. New York: Wiley, 2002.
34. Noble SM, Hollingworth W, Tilling K. Missing data in trial-based cost-effectiveness analysis: the current state of play. Health Econ 2012;21:187-200.
35. Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW. Missing data: a systematic review of how they are reported and handled. Epidemiology 2012;23:729-32.
36. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 2011;30:377-99.
37. Willan AR, Briggs AH, Hoch JS. Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. Health Econ 2004;13:461-75.
38. Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, Swinburn P, Busschbach J. Measurement properties of the EQ-5D-5L compared to EQ-5D-3L across eight patient groups: a multi country study. Qual Life Res 2012;22:1717-27.

Figure 1: Mean utility scores for treatment and control groups over 2 years (24 months)

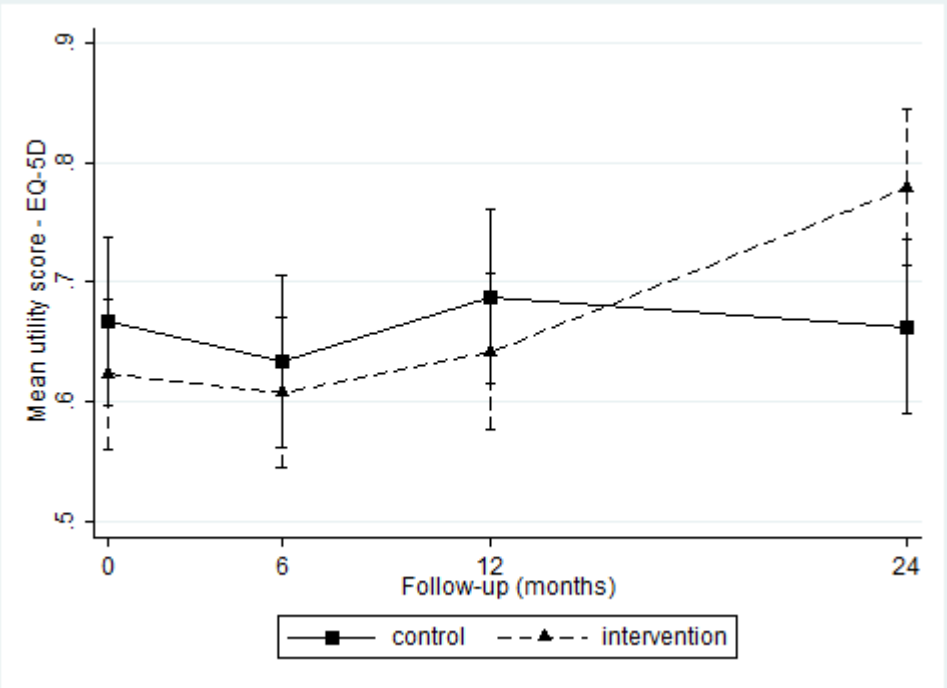

Figure 2: Frequency distribution of EQ-5D utility scores at baseline

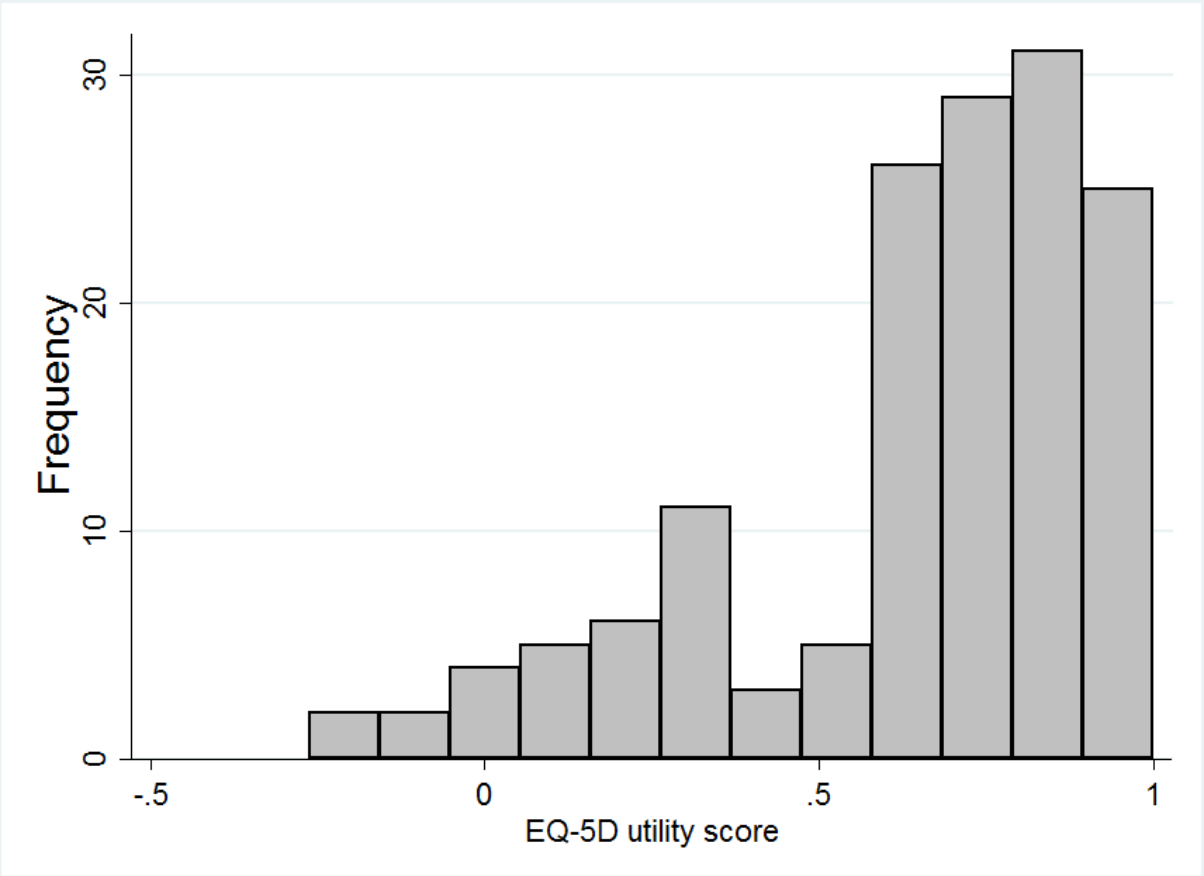

Table 1: Utility scores and QALYs by treatment group

	Mean (SD) Utility Scores		QALYS		Discounted	
Follow up point	Control	Treatment	Control	Treatment	Control	Treatment
Baseline	0.675 (0.28)	0.637 (0.3)				
6 months	0.631 (0.33)	0.605 (0.3)	0.327	0.31	0.327	0.31
12 months	0.70 (0.25)	0.644 (0.29)	0.333	0.312	0.333	0.312
24 months	0.711 (0.28)	0.805 (0.28)	0.71	0.724	0.682	0.70
Total	0.679 (0.28)	0.668 (0.30)	1.37	1.346	1.342	1.322

SD= standard deviation

Table 2: Percentage of individuals missing EQ-5D questionnaires or item responses

		Control	**Treatment**
Total – n		67	85
Baseline n (%)	Complete	66 (99%)	83 (98%)
	Missing 1 item on EQ-5D	1 (1%)	2 (2%)
	Missing EQ-5D	0	0
6 months n (%)	Complete	62 (92%)	82 (96%)
	Missing 1 item on EQ-5D	5 (8%)	2 (2%)
	Missing EQ-5D	0	1 (1%)
12 months n (%)	Complete	60 (90%)	72 (85%)
	Missing 1 item on EQ-5D	0	0
	Missing EQ-5D	7 (10%)	13 (15%)
24 months n (%)	Complete	57 (85%)	71 (83%)
	Missing 1 item on EQ-5D	4 (6%)	9 (11%)
	Missing EQ-5D	6 (9%)	5 (6%)
Complete across all time points		47 (70%)	60 (71%)

Table 3: Total QALYs over 24 months reported by method of analysis.

	Undiscounted		Discounted	
	Control	Treatment	Control	Treatment
Available Case				
Group Mean	1.37	1.35	1.34	1.32
Complete Case				
Group Mean	1.47 (1.36-1.57)	1.39 (1.27-1.51)	1.44 (1.34-1.54)	1.37 (1.25- 1.48)
Individual level	1.47 (1.36-1.57)	1.39 (1.27-1.51)	1.44 (1.34-1.54)	1.37 (1.25- 1.48)
Adjusting for baseline utilities	1.35 (1.28-1.43)	1.38 (1.32-1.45)	1.33 (1.26 -1.40)	1.36 (1.3-1.42)
Multi-level model	1.55 (1.47-1.62)	1.54 (1.44-1.56)	1.52 (1.44-1.60)	1.51 (1.42-1.63)
Missing data				
Last utility score carried forward	1.35 (1.22-1.46)	1.35 (1.25-1.46)	1.33 (1.21-1.44)	1.33 (1.22-1.43)
Last utility score carried forward (adjusted)*	1.32 (1.24-1.40)	1.37 (1.31-1.44)	1.30 (1.22-1.37)	1.35 (1.28-1.42)
Last item response carried forward	1.34 (1.23-1.46)	1.35 (1.25-1.46)	1.32 (1.20-1.44)	1.33 (1.22-1.44)
Last item response carried forward (adjusted)*	1.32 (1.24-1.39)	1.38 (1.31-1.44)	1.30 (1.22-1.37)	1.35 (1.28-1.42)
Multiple imputation (unadjusted)*	1.36 (1.25-1.47)	1.33 (1.24 –1.44)	1.33 (1.23-1.44)	1.32 (1.22-1.41)
Multiple imputation (adjusted)*	1.24 (1.15-1.33)	1.26 (1.19 –1.33)	1.22 (1.13 –1.30)	1.24 (1.16 –1.31)

*Adjusted: values adjusted for baseline utilities