Predicting missing quality of life data that were later recovered: an empirical comparison of approaches

Shona Fielding, Peter Fayers and Craig Ramsay

DOI: 10.1177/1740774510374626
Clin Trials 2010 7: 333 originally published online 24 June 2010
The online version of this article can be found at:
http://ctj.sagepub.com/content/7/4/333

Introduction

Missing data are a common problem in many areas of research. When a dataset includes a large proportion of missing data, subsequent analysis may be subject to bias unless dealt with appropriately. Possible approaches include complete case or available case analysis, joint or pattern mixture modelling and use of imputation [1]. Imputation is the substitution of an estimated value for one that is missing. There are two types of imputation – single value or multiple imputation.

Outcomes from health-related quality of life (QoL) instruments are considered. Missing QoL outcomes are likely to be informative [2,3] . For example, patients with poorer QoL may have impaired physical functioning and may feel fatigued, therefore may be less likely to complete questionnaires. Conversely those who feel well might regard the questionnaires as no longer relevant, thereby not filling them in. There are two types of missing data: missing form (the focus here) when the whole questionnaire is missing, and missing item, when the form is returned with one or more items unanswered.

There is extensive literature on the accuracy of different imputation strategies [4-12]. A number of case studies and simulation studies have shown multiple imputation (MI)

to be superior to simple imputation in the presence of informative missing data [4-12]. However in these studies the missing data have been created artificially and the missing data patterns are predetermined or prespecified. It is not surprising that some imputation methods perform poorly for particular prespecified mechanisms. For example, simple methods such as mean imputation are only useful if the data are likely missing completely at random (MCAR). If not MCAR, this type of method is likely to produce biased results. Fielding *et al.* [2] showed that simple imputation was inadequate in the presence of missing not at random (MNAR) data and suggested MI might be more appropriate.

The accuracy of different strategies of dealing with the missing data can never be truly assessed since the data are by definition 'missing'. However, this paper takes an approach that compares imputation results against data subsequently recovered using a reminder system for questionnaires. This allows the relative efficiency of different methods to be assessed. Use of imputation should always be regarded with circumspection, and we discuss the merits of the different approaches.

Dealing with missing data

Imputation

Imputation involves replacing a missing value with a 'best guess'. This could be based on previously observed scores (e.g. the last observation available for a particular person), expert opinion or previous studies. Additional data that may inform imputation are often contained within a dataset and can be utilised. The imputation process results in a complete dataset on which standard statistical analysis can be

carried out. However, imputation cannot completely replace lost information and should be used with caution. Although a seemingly complete dataset is created, it is one that has been augmented in some sense and does not compensate for full information [13]. Imputed values will never be completely representative of the true value.

Simple imputation

In simple imputation a single value is substituted for one that is missing. Methods consist of those using data from all patients (cross-sectional) and those using information specific to the patient (longitudinal). Many cross-sectional methods exist [13] but those used here are: simple mean imputation (mean score from observed data at a given assessment); minimum or maximum value of the observed data; and regression (using QoL and other variables in the dataset). For regression imputation the covariates associated with reminder response and the outcome can be identified using standard statistical tests such as independent t-tests and chi-squared tests as appropriate. A common but naïve longitudinal method is last observation carried forward (LOCF), where the last observed value is carried forward to replace a missing observation. This is sometimes referred to as last value carried forward. A similar method is baseline carried forward, where the baseline value is carried forward to each missing value.

Each of these simple imputation methods provides a deterministic rule for imputing a single value for each that is missing. The advantage of simple imputation is that it is easy to carry out. However, a major disadvantage, especially for methods based on mean values or regression, is that variances of the variable undergoing imputation

tend to become underestimated, leading to reduced standard errors which can impact on test statistics, p-values and confidence intervals [14].

Multiple imputation

MI can overcome some of the problems associated with simple imputation. For each missing value, instead of a single deterministic estimate, a random value is selected from its distribution. This introduces uncertainty in the imputed values, and preserves the random variability (variance) of the variables. This process is repeated, generating multiple randomly different datasets. Each dataset is analysed separately and the results combined using Rubin's method [15]. Although a single random imputation may be reasonable if the proportion of missing data is small, without special corrective measures the inferences tend to overstate precision because the between-imputation component of variability is omitted. Imputations may in principle be created under any kind of assumptions or model for the missing data mechanism and the resulting inferences will be valid under that mechanism. Several approaches were used for randomly sampling from the distribution of the items with missing values and these are described.

Approaches for monotone missingness

A monotone missing data pattern occurs when no further observations are made on a patient following a number of completed assessments. For monotone missing data possible MI methods include: regression models, predictive mean match models or propensity scores. For the regression method a model is fitted to the continuous outcome with explanatory variables constructed from those covariates available within the dataset. Based on the fitted regression model, a new regression model is

simulated from the posterior predictive distribution of the parameters and this is used to impute the missing values for each variable [16].

The predictive mean match model also fits a regression model and obtains a set of predicted values. For each piece of missing data, a set of observed values for which the predicted values are closest to the predicted value of the missing observation are obtained. One of these observed values is randomly selected and imputed for the missing value. An advantage of this method over regression is that imputed values are always within the range of the data and thus ensures the imputed values are plausible.

The third method for monotone missingness is the propensity score method. For a variable with missing values, a score is generated for each observation to estimate the probability that the observation is missing and this is called the propensity score. The observations are then grouped (usually five) according on these propensity scores. Within each group a set of observed scores are randomly selected with replacement to create a new set of observed scores. For each missing value an observation from this new set of observed scores is randomly selected and imputed. The disadvantage of this method is that information about correlations of repeated measures is lost. For all these methods formulation of the imputation model is an important step [17]. Failure to accommodate the model structure appropriately can cause bias in the resulting analysis [18]. Additional variables which are related to both missingness and outcome should also be included in the imputation model [1].

Intermittent missingness

Intermittent missing data occurs when one or more observations for a patient were missing before one was observed. For intermittent missing data, the Markov Chain

Monte Carlo (MCMC) method can be used. The exact detail behind MCMC is provided elsewhere [15-16, 19]. MCMC can be used in two ways for intermittent missingness. The first approach applies MCMC on all the intermittent missing data and the second uses MCMC to make the data monotone and then employs one of the other more flexible monotone methods for the remainder of the data.

Software for imputation

Simple imputation does not require any specialist software and the routines can be programmed into any standard statistical software. MI is more complex, however, and some statistical software packages provide inbuilt procedures. The statistical software package STATA undertakes MI by chained equations using the ICE command [20]. The software package SAS uses the multiple imputation procedure (PROC MI) to carry out the imputation followed by PROC MIANALYZE to combine the results [19]. The package, SOLAS (Statistical Solutions Inc, Sargus, MA, USA) has been developed to handle missing data and perform MI. The work presented here used SAS and full technical details of these imputation procedures can be found in the SAS User's Guide [19].

Repeated measures model

QoL data collected in trials is often analysed using an analysis of covariance (ANCOVA) on the final endpoint. Since longitudinal data has been collected an alternative is the repeated measures model [1,21]. This type of model assumes the data are MAR. It has several advantages over the simpler ANCOVA. It takes into account correlations between the repeated measures and allows for missing data [21].

The models are fitted using restricted maximum likelihood estimation, using for example the MIXED procedure in SAS [19].

Pattern mixture model

Pattern mixture models allow for different response models for each pattern of missing values [1,22]. The data are then a mixture of these weighted by the probability of each missing value or dropout pattern. To apply a pattern mixture model, the proportion of subjects for each pattern of missing data needs to be known. However, there can be a large number of potential patterns of missing data, causing difficulties in estimating all the model parameters in each pattern. Furthermore, for some patterns the model can be under identified, and not all the parameters can be estimated without additional assumptions or restrictions placed on the model [22]. Three possible restrictions for monotone missingness have been proposed: complete case missing value (CCMV); available case missing value (ACMV); neighbouring case missing value (NCMV) [1,15,22].

In summary under the CCMV restriction, the data from subjects in pattern one (complete cases) are used to impute the means for the missing observations in the remaining patterns. It is important to note that this restriction is only feasible when the number of cases in pattern one is sufficient to estimate these parameters reliably. In the ACMV restriction, available data from subjects in all patterns are used to impute the means for the missing observations in the remaining patterns. This is less restrictive than CCMV restriction as more observations are used to estimate some parameters. Finally, for the NCMV restriction available data from subjects in the neighbouring pattern are used to impute the means for the missing observations in the

remaining patterns. These three sets of restrictions result in a number of equations that need to be solved to obtain the unknown means and variance parameters. However, deriving the appropriate variance of the pooled estimates is complex. Curran [23] suggests an analytic technique using MI to avoid this problem and this has been implemented here.

Datasets

The datasets involved in this empirical work come from the Centre for Healthcare Randomised Trials in Aberdeen and include the physical and mental summary components of the SF36 [24] and the EuroQoL EQ5D score [25]. Data were collected at baseline (at a clinic appointment) with subsequent follow-up through postal questionnaires. To reduce the number of unreturned follow-up questionnaires, one or more reminders were issued to those who did not respond within a specified time (usually two weeks). This recovered a substantial portion of otherwise missing data.

Five randomised trials are presented. Several quality of life instruments were administered in the trials, but only the EQ5D results are presented in detail.

1. REFLUX (N=357) evaluated the clinical effectiveness and cost-effectiveness of early laparoscopic surgery compared with continued medical management amongst people with gastro-oesophageal reflux disease. QoL data (EQ5D, SF12 and gastro-oesophageal reflux symptoms) were collected at baseline, three months and twelve months after surgery (or equivalent for those managed medically). Trial analysis consisted of ANCOVA of the 12 month treatment difference (surgical versus medical management) adjusting for baseline QoL and a number of other baseline characteristics (age, sex and body mass index (BMI)) [26].

2. MAVIS (N=910) was a randomised controlled trial of multi-vitamin and mineral supplementation in persons aged 65 and over to reduce infection rates and antibiotic usage. QoL data (EQ5D and SF12) were collected at baseline, six and twelve months follow-up. ANCOVA was used to estimate the mean difference at 12 months between groups (placebo versus supplementation) after adjusting for baseline QoL values and the baseline covariates - age group, sex and type of housing [27].
3. RECORD (N=5292) was a placebo-controlled trial of daily oral vitamin D and calcium supplementation in the secondary prevention of osteoporosis related fractures in older people. QoL was assessed at 4, 12, 24, 36 and 48 months using the EQ5D and the SF12 instruments. An ANCOVA adjusting for baseline QoL, age group, gender, time since recruiting fracture and type of fracture was carried out to assess treatment effects on QoL at 24 months. The calcium supplementation (or no supplementation) treatment comparison will be used for illustration [28].
4. KAT aimed to measure the long-term clinical and cost effectiveness of different types of knee replacement. There were 1517 patients randomised between patella resurfacing or not. Functional status (Oxford Knee Score) and quality of life (SF12 and EQ5D) were measured at baseline, three months and annually after their operation. An ANCOVA for the two-year treatment comparison adjusting for baseline QoL, age group, sex and extent of arthritis was carried out [29].
5. PRISM (N=1324) evaluated the clinical effectiveness and cost-effectiveness of symptomatic versus intensive bisphosphonate therapy for the management of Paget's disease. QoL was assessed through yearly postal questionnaires including the EQ5D, SF36 and disease-specific QoL as measured by the Arthritis Specific Health Index. The primary endpoint in the PRISM trial was treatment effect at two

years adjusting for baseline QoL, and a number of Paget's disease-related variables [30].

The missingness mechanism in our studies was found to be a mixture of MCAR, missing at random (MAR) and possible MNAR [3]. Knowing this mechanism is important in determining which of the methods considered here are likely to provide the least biased results.

Methods

In this paper we use an empirical approach to compare the different strategies for dealing with missing data. The approach outlined in this paper differs from previous literature in that it is based on real data from trials in which initially missing data was later recovered using a reminder system. This paper deals with missing forms, so imputation is carried out on complete QoL dimensions rather than individual items comprising the dimension. It is only the data from the questionnaires that were obtained by reminder that are imputed and the not the data from the questionnaires that were never obtained. The data collected by reminder are initially regarded as missing, as if no reminder system had been used. This portion of data is then imputed and results from analyses of the trial compared to what was actually obtained when using all responders (including the reminder data). The impact of the imputation method on estimation of treatment difference is evaluated and the different strategies compared.

For each trial dataset, the relevant covariates for imputation were identified using standard statistical procedures (e.g. t-tests and chi-squared tests) to identify those that

were significantly associated with both the outcome (QoL) and the indicator of reminder response. Two imputation models were then used. Firstly a model including only covariates (involved in the original trial analysis plus any additional variables related to missingness and outcome) and secondly the same covariates plus previous QoL.

A repeated measures model was carried out for each trial. The baseline assessment was used as a covariate rather than incorporated into the repeated measures. The model also adjusted for the same covariates that were used in the original trial ANCOVA. This allowed the treatment difference estimates to be more comparable to the original analysis carried out by the trial researchers.

Determining the best method

For each dataset, the calculated treatment difference and its corresponding 95% confidence interval were obtained for the observed data, data under simple imputation, under MI, from the repeated measure model and from the pattern mixture models. The absolute bias in the calculated treatment effect was calculated. The full range of results is provided for the EQ5D QoL score and brief details given for the SF12 summary scores.

Secondly the precision of the estimate is important when determining accuracy of the different methods. The width of the confidence interval for the observed result and that under the imputation/modelling strategies was obtained. The ratio of this width was calculated. Ideally the ratio would be equal to one such that the observed precision was also seen in the imputation. The ‘best’ method was identified as the method which showed the smallest bias, but also took into account the precision. In an

ideal world a ‘perfect imputation’ would result in an estimate with no bias and variance equal to variance in the observed data and it is against this standard that the ratio of CI width was assessed.

Results

Table 1 shows the percentage of each type of responder from the total sample size (immediate, reminder or non-responder) at each assessment for each trial. It is seen that MAVIS had an excellent response rate to the initial mailing, while in REFLUX a large amount of data were recovered through the reminder responses.

In each trial analysis an ANCOVA model for treatment difference in QoL scores at the final endpoint adjusting for other covariates was carried out. This final endpoint was at one year follow-up for REFLUX and MAVIS, but at two years for RECORD, KAT and PRISM. The proportion of missing responses (or reminder-response here that will be imputed) from the total number of responders at these final endpoints was: REFLUX – 57%; MAVIS – 12%; RECORD – 22%, KAT – 18%; PRISM – 18%. Table 2 shows the covariates for each trial that were identified as being significant in the models for simple regression imputation and those involved in the MI procedures. Table 3 shows the observed treatment difference (95% CI) in EQ5D scores. The ‘absolute bias*100’ and ‘ratio of the CI width’ are presented in Table 3 for each of the imputation and modelling methods. The results under simple imputation, multiple imputation and the alternative procedures are discussed in turn. This is followed by a comparison between these different options for dealing with missing data.

Simple imputation

The choice of simple imputation method can have an impact on whether the calculated treatment difference is significant. For example in the REFLUX trial the methods of BCF, LOCF and regression all provide a significant treatment difference ($p<0.05$). In the REFLUX and MAVIS trials the baseline carried forward (BCF) method provided the smallest total bias of the simple imputation procedures considered. The precision based on the ratio of CI width was equal to one in MAVIS using BCF, but in REFLUX this ideal value occurred with LOCF. In RECORD and KAT, LOCF provided the least biased estimates of treatment difference (bias = 0.2 and bias = 0. 4 respectively) and for RECORD the ratio of CI width was equal to one. Mean imputation and regression imputation tended to provide most bias. In the PRISM trial the maximum method showed least bias (bias = 0.4) compared to 1.0 for the BCF method. However, BCF was much better at maintaining precision as the ratio of CI width was one.

There were 13 other QoL scores measured across the five trials in addition to EQ5D. LOCF was the most accurate (least biased) in five of these. BCF showed least bias on four occasions, mean value imputation for three QoL scores and finally maximum value imputation for one instrument. Combining these results with the EQ5D data shows that the longitudinal simple imputation methods (BCF and LOCF) provided the greatest number of 'best' (least biased) estimates (13 of 18 QoL scores).

Multiple imputation

As with simple imputation the choice of MI had an impact on the calculated estimates of treatment difference. MCMC imputation was carried out on all the missing data as it allows for the intermittent missing data pattern. For the regression, predictive mean

match and propensity score methods, MCMC was first used to make the data monotone. The predictive mean match model was the least biased method for four of the five trials (REFLUX, MAVIS, RECORD and PRISM). MCMC for intermittent missingness showed an equivalently small bias as the predictive mean match for the RECORD trial. In KAT, the regression model (including previous QoL scores) resulted in least bias. The ratio of CI width was reasonable (equal to or close to one) for the predictive man match model in each of the trials.

MI for the other 13 QoL scores within the trials showed that MCMC for intermittent missingness was least biased for four of them, as was the predictive mean match model. A propensity model was least biased on three occasions and finally regression was the least biased MI method for only two of the twelve QoL scores.

Other strategies

Pattern mixture models (incorporating MI) did not perform that well for REFLUX or KAT data in terms of bias. In the MAVIS trial under CCMV the calculated treatment estimate showed a much greater bias than the other two restrictions and the ratio of CI width was much larger than one. Across all trials, of the three restrictions it was either ACMV or NCMV which performed best with NCMV tending to provide the better precision. The estimates obtained using a repeated measures model were more accurate than pattern mixture modelling for the REFLUX, MAVIS, RECORD and KAT trials. Pattern mixture models appeared to perform better for PRISM in terms of bias, although the precision was reasonable in both cases. However one note of caution is that the results for the pattern mixture models are based on only those patients with a monotone missingness pattern.

Comparison between the different approaches for dealing with missing data

The paragraphs above have discussed in turn the results of the simple imputation, MI, repeated measures model and pattern mixture models. Table 3 shows the calculated bias (bias*100) for each of these strategies for the EQ5D score in the five different trials. In four of the five trials, one of the MI strategies showed the least bias. In the fifth a pattern mixture model on the monotone missing data (using NCMV, bias = 0.1) was least biased but this was closely followed by two of the MI procedures (bias = 0.2) carried out on all the missing data. The most accurate MI strategy was the predictive mean match model. On the whole it was one of the MI strategies which provided the best precision – that is, the ratio of the CI width was either equal or close to one. In these five trials MI consistently provided results with low bias and best precision.

Discussion

It has been shown in this paper that different imputation strategies can impact on the eventual calculated treatment estimates to varying degrees. Generally the longitudinal methods (BCF and LOCF) were the 'best' simple imputation methods. MCMC imputation to make the data monotone followed by a predictive mean match model was good and the best MI method for the EQ5D scores in four of the trials. Usually one of the MI procedures provided least biased results and the most precise. This was true for the data presented and for the other QoL measures used within the five trials. The ratio of CI width was close to one and the standard errors tended to reflect those calculated in the observed dataset. This is important when calculating the confidence intervals and p-values for significance [14]. The MI strategies seemed preferable to the other procedures considered.

The REFLUX trial contained the most missing data (and correspondingly, the greatest amount of data collected via reminder-response). MCMC to make the data monotone followed by a predictive mean match model was clearly more superior over the simple imputation methods. Pattern mixture models for this dataset performed particularly badly when compared to the standard MI models or the repeated measures model. The difference between simple and MI is less obvious as the amount of data being imputed is reduced. For example, in the MAVIS trial only 12% of data is undergoing imputation and in this trial one of each of the simple and MI methods were equivalent in providing the smallest bias.

Previous work has shown evidence against MCAR for the five datasets [3]. Therefore, it would be unlikely that simple imputation would be appropriate, whereas MI might be more suitable. The simple imputation methods assume that data are MCAR – unrelated to anything observed. MI methods assume MAR and that missing data are related to observed data (covariates and/or outcome) [1]. If there is no evidence for MCAR then simple imputation methods should be used with caution or not at all. It is likely that QoL data is frequently missing for a reason related to changing QoL and so MI should be preferable because MAR is more plausible. The CCMV restriction for the pattern mixture model assumes MCAR and therefore not surprisingly this performs worst of the three restrictions. ACMV and NCMV are based on the MAR assumption which as we have already discussed is more plausible in this context. The pattern mixture models were carried out on a reduced dataset which only contained those patients with a monotone missing data pattern. This may account for the differences seen between ACMV/NCMV and the MI models.

Molenberghs and Kenward [14] discuss the merits of the different MI approaches. They promote the use of a regression or predictive mean match model for the longitudinal setting. An advantage of a predictive mean match model over regression is that imputed values are always within the range of the data [1]. In situations of monotone missingness, it is expected that the MCMC approach and the regression method should lead to similar answers. Any difference is due largely to the different prior distribution used [14]. Regression and predictive mean match imputation have been shown to be the most accurate in this current situation, with the propensity score model on the whole performing poorly in comparison. A reason for this is given by Molenberghs and Kenward [14] (page 144):

> "The propensity score method uses only the covariate information associated with whether the imputed values are missing. It does not use associations among variables; As a consequence, while it can be effective for inferences about the distributions of individual imputed variables, it is not appropriate for analyses involving relationships among variables." [14]

If the model used for analysis and the imputation model are the same, the resulting estimates under imputation will be equivalent to those obtained by maximum likelihood, for example using a repeated measure design [17]. However, when the imputation model uses covariate information which is potentially related to the missing QoL values, the estimates may be different.

There is increasing evidence from simulation studies that MI provides more robust treatment estimates and appropriate standard errors [4-12]. Each of these

authors has compared MI to a simpler alternative or complete case strategy. In some situations MI was shown to perform at least as well in terms of treatment effects but on the whole provided a more realistic standard error. Our study complements these studies using real data with reminders rather than simulated missing data.

One note of caution of the use of imputation is with regard to its validity when there is a large proportion of missing data. For example, the proportion of data obtained through reminders for REFLUX was 57% and to impute this much data should be done so with caution. Imputation is often regarded as a tool to assess sensitivity of results to missing data rather than a primary analysis [1,4,31]. The methods presented here could be used to identify which methods would be most appropriate. One limitation of the current work is that we have not considered the impact of the different strategies on the eventual trial result. This is an important consideration and is addressed in a future paper.

A second limitation was the use of the criteria of the ratio of CI width being equal to one as a proxy for precision of the imputation procedure. A perfect imputation procedure would result in the variance of the imputed estimate being equal to the variance of the full data estimate and the ratio equal to one. However, in practice it might be that the value of one was obtained due to a bias in the imputed estimate and thus it is possible that a value of one does not suggest an optimal imputation. In situations where there is a large amount of missing data, the variance is likely to be underestimated and the criteria of the ratio of CI width has less standing. However, since it is unlikely imputation

would be used in these situations anyway, despite the concern, we felt that using this ratio of CI width as a measure of precision was a fair assumption to make.

The rationale underlying our approach is that the 'reminder-responders' are likely to be representative of those who do not respond at all. Thus we were able to identify possible suitable imputation methods or an appropriate modelling alternative. This method could then be used on the actual missing data to allow more patients (and data) to be included in the final analysis. Although it can never be proved that they are representative, it was found that the mechanism behind reminder-response was usually the same as the mechanism behind non-response for each of these trials [3]. This suggests that the rationale behind the approach presented in this paper is valid. However since the data required to prove this are missing, the strategies outlined should be used as a sensitivity to results rather than a primary analysis [17].

Conclusion

As Huson *et al.* [7] observe there is no one imputation technique that is applicable for all possible missing data patterns and missing data mechanisms. However, here we have shown that MI was more suitable then both simple imputation methods and repeated measures models when missing data was found to be informative. MI models the uncertainty in the missing data and is based on the MAR assumption, which is more plausible in the QoL setting. When deciding on the best model for imputation it is recommended that all the variables in the analysis model are included plus any additional variables which are related to both outcome and missingness. MI is the hard way to analyze data where missingness is MAR and will only provide a benefit

when the analyst has additional information that is related to QoL both when the response is observed and when it is missing [1]. We suggest that where possible reminder data should always be collected and can be used to identify suitable imputation procedures. This 'best' choice can then be used on the actual missing data to allow more patients to be included in analysis.

Abbreviations

ACMA – available case missing value; ANCOVA – analysis of covariance; BCF – baseline carried forwards; BMI – body mass index; CCMV – complete case missing value; CI – confidence interval; EQ5D – EuroQoL EQ5D; ICE – imputation by chained equations; KAT – Knee Arthroplasty Trial LOCF – Last observation carried forward; MAR – missing at random; MAVIS – Randomised trial of mineral and vitamin supplementation; MCAR – missing completely at random; MCMC- Markov Chain Monte Carlo; MI – multiple imputation; MNAR – missing not at random; NCMV - neighbouring case missing value; NVCB – next value carried backwards; PRISM – Paget's disease: a Randomised trial of Intensive versus Symptomatic Management; RECORD – Randomised Evaluation of Calcium and/OR vitamin D; REFLUX- Randomised Evaluation oF Laproscopic sUrgery for reflux; SF12/36 – Short Form 12 or Short Form 36; QoL – Quality of life; RCT – Randomised controlled trial.

Acknowledgments

We would like to thank the Centre for HealthCare Randomised Trials based within the Health Services Research Unit and their staff for providing the data used for this work. Particularly, Gladys McPherson, Alison McDonald, Graeme MacLennan,

Jonathan Cook and Samantha Wileman who assisted with data queries and provided background to the trials.

Funding

The Health Services Research Unit is funded by the Chief Scientist Office of the Scottish Government Health Directorate. Shona Fielding was funded by the Chief Scientist Office on a Research Training Fellowship (CZF/1/31) while carrying out this work. The views expressed are, however, not necessarily those of the funding body.

References

(1) Fairclough DL. Design and Analysis of Quality of Life Studies in Clinical Trials. : Chapman and Hall; 2002.

(2) Fielding S, Fayers PM, McDonald A et al. Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. Health & Quality of Life Outcomes 2008; 6 (57).

(3) Fielding S, Fayers PM, Ramsay CR. Investigating the missingness mechanism in quality of life data: A comparison of approaches. Health & Quality of Life Outcomes 2009; 7(57).

(4) Cook NR. An imputation method for non-ignorable missing data in studies of blood pressure. Statistics in Medicine 1997 ; 16(23): 2713-2728.

(5) Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J.Clin.Epidemiol. 2006; 59(10):1087-1091.

(6) Hunsberger S, Murray D, Davis CE, Fabsitz RR. Imputation strategies for missing data in a school-based multi-centre study: the Pathways study. Statistics in Medicine 2001; 20(2):305-316.

(7) Huson LW, Chung J, Salgo M. Missing data imputation in two phase III trials treating HIV1 infection. Journal of Biopharmaceutical Statistics 2007;1 7:159-172.

(8) Liu G, Gould AL. Comparison of alternative strategies for analysis of longitudinal trials with dropouts. J.Biopharm.Stat. 2002; 12(2):207-226.

(9) Morita S, Kobayashi K, Eguchi K, Matsumoto T, Shibuya M, Yamaji Y, et al. Analysis of incomplete quality of life data in advanced stage cancer: A practical application of multiple imputation. Quality of Life Research 2005; 14(6):1533-1544.

(10) Myers WR. Handling missing data in clinical trials: An overview. Drug Inf.J. 2000; 34(2):525-533.

(11) Patrician PA. Multiple imputation for missing data. Res. Nurs. Health 2002; 25(1): 76-84.

(12) Tang L, Song J, Belin TR, Unutzer J. A comparison of imputation methods in a longitudinal randomized clinical trial. Stat.Med. 2005; 24(14): 2111-2128.

(13) Fayers PM, Machin D. Quality of Life: Assessment, Analysis and Interpretation. Wiley, 2001.

(14) Molenberghs G, Kenward MG. Missing Data in Clinical Studies. Wiley, 2007.

(15) Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed; Wiley, 2002.

(16) Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, Inc; 1987.

(17) Carpenter JR, Kenward MG. *Missing data in randomised controlled trials - a practical guide*. November 2007 Available: http://www.pcpoh.bham.ac.uk/publichealth/methodology/docs/invitations/Final_Report_RM04_JH17_mk.pdf [2007, 28/11].

(18) When are inferences from multiple imputations valid? Proceedings of the Survey Research Methods Section of the American Statistical Association; 1992.

(19) SAS Institute Inc. SAS/STAT 9.1 User's Guide. 2004.

(20) Royston P. Multiple imputation of missing values: update of **ice**. Stata Journal 2005; 5:527.

(21) Brown H, Prescott R. Applied Mixed Models in Medicine. Wiley, 1999.

(22) Little RJA. A class of pattern-mixture models for normal incomplete data. Biometrika 1994; 81(3):471-483.

(23) Curran D, Molenberghs G, Thijs H, Verbeke G. Sensitivity analysis for pattern mixture models. J.Biopharm.Stat. 2004;14(1):125-143.

(24) Ware JR, Snow KK, Kosinski M., Gandek B. SF-36 Health Survey Manual and Interpretation Guide. 1993.

(25) Brooks, R with the EuroQoL Group. EuroQoL: the current state of play. Health Policy 1996; 37:53-72.

(26) Grant A, Wileman SM, Ramsay C et al. The effectiveness and cost-effectiveness of minimal access surgery amongst people with gastro-oesophageal reflux disease - a UK collaborative study. The REFLUX trial. Health Technology Assessment 2008; 12(31):1-204.

(27) Avenell A, Campbell MK, Cook JA et al. Effect of multivitamin and multimineral supplements on morbidity from infections in older people (MAVIS trial): pragmatic, randomised, double blind, placebo controlled trial. BMJ 2005; 331(7512): 324-329.

(28) Grant AM, Avenell A, Campbell MK et al. Oral vitamin D3 and calcium for secondary prevention of low-trauma fractures in elderly people (Randomised Evaluation of Calcium Or vitamin D, RECORD): a randomised placebo-controlled trial. Lancet 2005; 365(9471):1621-1628.

(29) The KAT trial group. The Knee Arthroplasty Trial (KAT) Design Features, Baseline Characteristics and Two-Year Functional Outcomes after Alternative Approaches to Knee Replacement. J Bone Joint Surg Am 2009(91):134-141.

(30) Langston AL, Campbell MK, Fraser WD et al. Randomised Trial of Intensive Bisphosphonate Treatment Versus Symptomatic Management in Paget's Disease of Bone. Journal of Bone and Mineral Research (doi: 10.1359/jbmr.090709) (in press).

(31) Fayers PM, Machin D. Quality of Life: The assessment, analysis and interpretation of patient-reported outcomes. 2nd ed. UK: Wiley; 2007.

Tables

Table 1: Percentage of each type of responder in each trial

		Type of responder (%)		
Trial	Assessment	Immediate	Reminder	Non-responder
REFLUX (N=357)	3 months	39	47	14
	12 months	38	51	11
MAVIS (N=910)	6 months	91	4	5
	12 months	81	11	8
RECORD (N=5292)	4 months	58	20	22
	12 months	54	17	29
	24 months	51	14	35
KAT (N=2356)	3 months	79	9	12
	1 year	74	13	13
	2 years	69	15	16
PRISM (N=1324)	1 year	85	6	9
	2 years	63	14	23

Table 2: Covariates involved in the imputation models

Trial	Covariates involved in simple imputation regression	Covariates involved in multiple imputation models
REFLUX	Gender, age group, BMI group, treatment, baseline QoL	Treatment baseline QoL sex, age, BMI
MAVIS	Gender, residence type, age group, baseline QoL	Age group, gender, residence type, treatment, presence of chronic infection
RECORD	Gender, time since recruiting fracture, fracture type, age group, 4m QoL	Treatment, gender, age group, time since recruiting fracture, type of fracture (proximal or distal or vertebral), age group, residence type after fracture
KAT	Gender, age, ASA grade	Age group, treatment, extent of knee arthritis, any hospital readmissions, further knee admissions
PRISM	Baseline QoL, Treatment, a number of Paget related variables	Treatment, age and a number of Paget's related variables

Table 3: Accuracy results of estimates using different approaches – Bias*100 (Ratio of CI width)

	REFLUX	MAVIS	RECORD	KAT	PRISM
% of responses from reminder	*57%*	*12%*	*22%*	*18%*	*18%*
Observed treatment difference (Units*100)	*4.7*	*-1.9*	*1.5*	*1.3*	*1.5*
Observed 95% CI (units*100)	*(-0.3,10)*	*(-4,0.2)*	*(0,30)*	*(-1.0,4.0)*	*(-2.0,5.0)*
Simple imputation methods	**Bias*100 (Ratio of CI width)**				
Mean	3.1 (0.7)	0.6 (1.0)	0.5 (0.8)	0.9 (1.2)	1.2 (1.3)
Maximum	1.9 (0.8)	1.2 (1.2)	1.0 (1.0)	1.4 (1.2)	0.4 (0.9)
Baseline carried forward	1.4 (0.6) †	0.2 (1.0)	0.4 (0.7)	0.8 (1.0)	1.0 (1.0)
Last observation carried forward	2.1 (1.0) †	0.8 (1.0)	0.2 (1.0)	0.4 (1.2)	0.9 (0.9)
Regression	1.8 (0.5) †	0.4 (0.8)	0.4 (0.9)	0.8 (1.0)	1.1 (0.4)
Multiple imputation methods					
MCMC for intermittent	4.5 (1.4) †	0.7 (1.2)	0.2 (1.1)	0.2 (1.2)	0.9 (1.0)
Regression model*	1.3 (1.6)	0.7 (1.4)	0.3 (1.2)	0.8 (1.0)	1.0 (1.0)
Predictive mean match model*	0.8 (1.1)	0.1 (1.0)	0.2 (1.2)	0.9 (1.0)	0.3 (1.3)
Propensity model*	1.6 (1.0)	0.4 (1.2)	0.4 (1.4)	0.6 (1.6)	1.1 (1.0)
Regression model**	2.6 (1.7)	0.6 (1.2)	0.4 (0.9) †	0.1 (1.2)	1.3 (1.0)
Predictive mean match model**	2.7 (1.7)	0.2 (1.0)	0.3 (1.3) †	0.3 (1.2)	0.6 (0.9)
Propensity model**	1.1 (1.4)	0.4 (1.2)	0.3 (1.3) †	0.9 (1.0)	0.9 (1.0)
Modelling strategies					
Repeated measures model#	3.3 (1.2) †	0.4 (1.2)	0.1 (1.1)	0.4 (1.2)	0.9 (1.0)
Pattern mixture (CCMV)	8.3 (1.4) †	19.9 (18.6)	0.4 (1.2)	1.8 (1.4)	0.6 (1.1)
Pattern mixture (ACMV)	8.3 (1.3) †	0.6 (1.2)	0.3 (1.2)	1.9 (1.4)	0.7 (1.0)
Pattern mixture (NCMV)	9.3 (1.2) †	0.6 (1.2)	0.1 (1.1)	1.9 (1.4)	0.6 (1.1)

* MI model based on the ANCOVA model and additional covariates; ** MI model based on the ANCOVA model, additional covariates and previous QoL;
# adjusted for covariates in original ANCOVA model; † $p<0.05$;
MCMC - Monte Carlo Markov Chain; CCMV- complete case missing value restriction;
ACMV - available case missing value restriction; NCMV - neighbouring case missing value
Bias = |observed treatment difference – treatment difference under imputation|
Ratio of CI width = width of 95% CI for treatment difference under imputation (or modelling) / width of 95% CI for observed treatment difference