This is the final draft, after peer-review, of a manuscript published in Contemporary Clinical Trials 2012;33(3):461-469. The definitive version, detailed above, is available online at www.sciencedirect.com

Analysing randomised controlled trials with missing data: choice of approach affects conclusions

Running Head: RCTs: choice of approach changes conclusions

Word Count: 4047

Dr Shona Fielding[1], Professor Peter Fayers[2,3], Dr Craig R Ramsay[4]

[1] Medical Statistics Team, Division of Applied Health Sciences, University of Aberdeen, UK

[2] Emeritus Professor, Division of Applied Health Sciences, University of Aberdeen, UK

[3] Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

[4] Health Services Research Unit, Division of Applied Health Sciences, University of Aberdeen, UK

Corresponding author:

Dr Shona Fielding

Medical Statistics Team

Division of Applied Health Sciences

University of Aberdeen

Polwarth Building

Foresterhill

Aberdeen

AB25 2ZD

Tel: +44 (0)1224 437107

Email: s.fielding@abdn.ac.uk

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.

ABSTRACT

Background

The publication of a wrong conclusion from a randomised trial could have disastrous consequences. Missing data are unavoidable in most studies, but ignoring the problem may introduce bias to the results. Finding an appropriate way to deal with missing data is of paramount importance. We show how the choice of analysis method can impact on the conclusion of the trial with regard to the quality of life outcomes.

Methods

Various analysis strategies (analysis of covariance, linear mixed effects model) with and without imputation were carried out to assess treatment difference in four quality of life outcomes in an example clinical trial.

Results

Across all four quality of life outcomes, the various analysis approaches provided different estimates of treatment difference, with varying precision, using different numbers of patients. In some cases the decision about statistical significance differed. The results suggested that where possible extra effort should be made to retrieve missing responses. In the presence of data missing at random, simple imputation was inappropriate with multiple imputation or a linear mixed effects model more useful.

Conclusion

Different trial conclusions were obtained for a variety of analysis approaches for the same outcome. Collecting as much data as possible is of paramount importance. Careful consideration should be taken when deciding on the most appropriate strategy for analysis when missing data are involved and this strategy should be pre-specified in the trial protocol. Making inappropriate decisions could result in inappropriate conclusions potentially leading to the adoption of a clinical intervention in error.

Keywords: clinical trial, analysis, missing data, quality of life, imputation

Introduction

The randomised controlled trial (RCT) is an important way of evaluating healthcare interventions, forming the basis of evidence based medicine [1]. Information gained from trials is optimal when the trial dataset is complete or relatively few data are missing. In practice this is very difficult to achieve and most trial datasets will contain missing data. Missing data are a problem for many different types of outcomes. Ignoring the presence of missing data could have major consequences and potentially lead to the publication of a wrong conclusion about a particular therapy, which ultimately could impact on clinical practice. Follow-up outcome data collected through postal questionnaires are particularly susceptible to the problems of missing data as completion cannot be enforced.

The focus of the work presented is quality of life (QoL) outcomes, but the results are applicable to the problem of missing data in general. Taking account of missing QoL outcome data is of paramount importance as often the reason why the data are missing is related to the QoL itself. Patients may forget to fill them in and not return the questionnaires, may not be physically or mentally able or perhaps do not receive them through being lost in the post. The missing data mechanism describes the underlying reason why missing data have occurred [2]. If missingness relates to the QoL itself then this could potentially be important when analysing the trial outcomes. If missingness is due to death the implications of this should be considered.

In an effort to tackle the problem of non-returned postal questionnaires some organisations now employ a system of reminder questionnaires to help retrieve data that were initially missing. The rationale being that sometimes participants need a little prompting and receiving a reminder may prompt them to respond, improving the sample size, allowing the

study to have sufficient power to make conclusions and not introduce bias through some participants being removed from analysis.

A common approach in analysing RCTs is to use a complete case analysis, whereby patients with incomplete data are ignored. In recent years the use of imputation has been seen as a way of providing a sensitivity analysis for this. Choice between imputation methods is often limited to those which are readily available and easy to implement (e.g. mean imputation). Recent advances in multiple imputation have caused this to be more widely used, but this approach is still considered as a bit of a 'black box' by many researchers [3]. Many trials (including our example, REFLUX) collect QoL outcome data at baseline and several times during follow-up, but only data from the final endpoint are analysed. A complete case analysis on the final endpoint ignores any patient without this final outcome even though their interim responses may be valuable in deciding between treatment options. Using an example trial we aim to investigate the use of alternative analysis strategies that utilise all responses and alongside different approaches for dealing with missing data show how conclusions about which treatment is best can be affected.

Methods

Example trial

The REFLUX trial [4, 5] was undertaken by Centre for Healthcare Randomised Trials, part of the Health Services Research Unit at the University of Aberdeen. The aim of this trial was to determine the relative benefits and risks of laparoscopic fundoplication surgery as an alternative to long term drug treatment for gastro-oesophageal reflux disease (GORD). It was a multicentre trial and recruited 357 participants (178 to surgery and 179 to medical management) to the randomised part of the trial and 453 to the preference arms. Since we are focusing on RCTs the analysis presented throughout relates to the 357 patients

recruited to the randomised arms of the trial. The primary outcome was the disease specific REFLUX quality of life score with the generic measures of QoL, SF36 and EQ5D as secondary outcomes. The REFLUX score ranges from 0 to 100 and was derived from the weighted average of six questions covering heartburn, acid reflux, eating and swallowing, bowel movements, sleep, and work, physical and social activities [6]. The SF36 provided two summary measures – physical summary and mental summary, each measured on 0-100 [7]. The EQ5D consists of five questions each with a three category response scale, resulting in 243 possible health states which are represented by a continuous outcome ranging from -0.59 (QoL worse than death) to 1 (best QoL) [8]. For each QoL outcome a higher score represents better QoL. The outcomes were assessed at baseline in clinical appointment and then via a postal questionnaire at follow up of three and 12 months post surgery or at an equivalent time for those being medically managed.

At each follow-up if a participant did not return the questionnaire within two weeks a reminder was issued and subsequently a second reminder two weeks later of they had still not responded. This generated an extra portion of data that would otherwise have been missing. This feature allows us to investigate the impact of the reminder strategy on the trial conclusion. Statistical analysis of QoL outcomes involved an analysis of covariance (ANCOVA) of the 12 month score adjusting for age, body mass index (BMI), sex and when appropriate baseline score and interaction between baseline score and treatment. This approach of analysing the final endpoint, ignoring any interim follow-up data has been found to be quite common [9].

Pattern and mechanism of missing data

Missing data occurs in one of two ways: missing items where one or more questions are missed from a returned questionnaire and missing forms where the whole questionnaire is

not returned. Many validated QoL questionnaires now allow for some missing items and scoring algorithms take account of this [7, 10]. This paper deals with the issue of missing forms. Within a study with multiple follow-up assessments, participants will display a pattern of missing data. If they return all questionnaires they are regarded as providing complete data. If they return all questionnaires until a time at which they fail to return anymore they display a monotone missing data pattern. An intermittent pattern of missing data occurs if a questionnaire is missed but the participant subsequently returns one at a later follow-up.

To understand how best to deal with missing data the first step is to determine the missing data mechanism. Three mechanisms of missing data were originally presented by Rubin: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [2]. MCAR represents the situation where the probability that an observation is missing does not depend on observed or unobserved data. MAR occurs when given the observed data, the probability that an observation is missing does not depend on unobserved data. MNAR means that after accounting for the available observed data, the reason for data being missing still depends on the unseen observations. It is usually impossible to prove that data are or are not MNAR as the data required to establish this are by definition missing and unknown [11]. In the context of QoL data, the mechanism refers to whether the missingness is somehow related to the QoL [11]. MCAR occurs if the reason for missing QoL assessment is entirely unrelated to QoL (e.g. the patient moved). When missingness is related to observed QoL (e.g. a previous assessment) after conditioning on covariates then the data are MAR. MNAR occurs if missingness is dependent on the QoL at the time the assessment is missing as well as on covariates and observed data. Previous work has shown that for REFLUX missing data were MCAR or potentially MAR [12]. This work utilised the data collected by reminder, pretended it were

missing and assessed the missing data mechanism. Undertaking this investigation in this way meant it was possible to determine if the reminder data were MNAR (as we did in fact know the observed data). Based on this we were able to conclude that potentially the actual missing data were most likely MCAR or MAR. This is an important finding when thinking about possible options for imputation or analysis, but as previously intimated we still cannot strictly rule out MNAR (for the actual missing data, rather than reminder data) as the actual data needed to do so are in fact missing. Perhaps more importantly the reason for being missing may depend on unknown factors, information on which is my definition not available.

Dealing with missing data

There are a number of possible methods which could be used to analyse the QoL outcomes to determine if there was a treatment difference. The REFLUX trial group used a complete case analysis and only used those patients for which baseline and 12 month scores were available. This approach utilised only 276 patients (77%) of the total recruited to the trial and ignored the data collected at three months. A further 29 participants provided either baseline or both baseline and three month data which could have been utilised. In total 353/357 (98.9%) participants provided a REFLUX score on at least one of the three occasions and all of these participants could be included if a repeated measures approach was used. Imputation has the potential to provide a value for each piece of missing data depending on the method used. For example, last value carried forward (LVCF) will only provide data for all patients if there is at least one value from a previous assessment available. This method could not be used for example at three months if the baseline values were missing. The more flexible multiple imputation method can potentially overcome all missing values depending on what variables are entered into the imputation model. The use of covariates which have missing values themselves can limit this process. Thus a number

of things have to be considered when deciding on an imputation model: the assumptions they make, the data they require and whether it is of benefit to only be able to impute a subset of the values that are missing. A number of options for dealing with missing data are now described.

(i) Complete case analysis

The easiest but usually least desirable option is to simply ignore the missing data and carry out a complete case analysis. This potentially removes a large number of people from the analysis and is likely to provide a biased result unless the mechanism is MCAR. In this paper this approach is implemented firstly on those responses received without reminder (referred to as immediate responses). Secondly, those responses collected after the participant had been issued with a reminder (referred to as reminder responses) can be included to provide a larger number of observations for analysis.

(ii) Imputation

A second option often considered is the use of imputation, whereby a reasonable alternative value is substituted in for one that is missing. Imputation can occur with a single value (simple imputation) or with multiple values (multiple imputation). Following imputation an augmented complete dataset is obtained, on which standard statistical procedures can be carried out. Common simple imputation methods are simple mean imputation, LVCF, hot-deck (random selection from observed responses) or regression [13].

The problems associated with simple imputation are well documented [3, 11, 13]. The majority of methods assume MCAR and will often underestimate the variances, resulting in inappropriate standard errors leading to inappropriate confidence intervals and p-values

[3]. Multiple imputation (MI) aims to overcome this problem and the recent developments in software mean that it is more readily available to the researcher [14]. MI techniques take account of the uncertainty surrounding the missing value and rather than a single value imputed, a number of imputations are carried out creating several augmented datasets. Each dataset is analysed separately and then the results combined using Rubin's method [15].

Several methods exist for multiple imputation and some require monotone missingness (when a participant drops out the study and provides no further assessments following a period of observed assessments). Regression, predictive mean match or propensity scoring can be used [14, 15]. If an intermittent pattern of missingness exists then Markov Chain Monte Carlo (MCMC) imputation using an approximate Bayesian bootstrap can be used [14]. Further details on multiple imputation can be found elsewhere [14, 15].

(iii) Model-based strategies

In the context of longitudinal data it is possible to use model-based strategies such as a mixed-effects model to deal with the missing data [17]. This type of modelling assumes MAR which is more plausible in the setting of QoL. More complex procedures exist such as a selection model, joint mixed effects model and pattern mixture models and these can account for MNAR if implemented carefully [11]. All of these methods require strong assumptions and these assumptions cannot formally be tested. The model-based strategy considered here was that of a linear mixed effects model. This allowed the interim information from intermediary assessments to be included. This approach increases the number of participants used in analysis as each can be included if they provide at least one QoL assessment (and it does not have to be the assessment of interest).

Methods implemented for REFLUX

Additional data in the REFLUX trial were collected via a reminder system as previously described. Not all researchers use such a system so although we have this data for REFLUX we will illustrate the use of some methods which would only use the immediate responses (with a view to showing having as much data as possible is of benefit). An analysis of covariance (ANCOVA) on the final endpoint is applied, but also the use of a linear mixed effects model which incorporates the three month data is explored. One example of a simple imputation method (LVCF) is applied alongside multiple imputation using a predictive mean match model [14]. These approaches were chosen on the basis of previous work [18]. In this previous paper a more comprehensive investigation into suitable imputation methods was reported.

The methods applied here are as follows:

1. ANCOVA at 12 months on immediate response data (no reminder responses)
2. ANCOVA at 12 months on all observed responses (immediate and reminder) – the published analysis approach
3. Last value carried forwards on immediate responses followed by ANCOVA
4. Last value carried forwards on all observed responses followed by ANCOVA
5. Linear mixed effects model on immediate responses only at three and 12 months
6. Linear mixed effects model using all observed responses
7. Predictive mean match MI model on immediate responses
8. Predictive mean match MI model on observed responses

Results

The REFLUX trial included 357 randomised participants. At the final endpoint (12 months), 38% responded immediately and a further 51% responded after reminder. This gave an overall response rate of 89%. The patient characteristics collected at baseline are shown in Table 1. The mean age was 46.3 years and two thirds were male. No obvious differences were seen between the two groups which was to be expected since the groups were randomised. Table 2 shows the missing data pattern for the REFLUX trial. Just over 80% of participants returned all three questionnaires.

Where appropriate imputation was carried out, after which each analysis approach outlined earlier was implemented for each of the four QoL outcomes. The estimate of treatment difference and its 95% confidence interval (CI) are presented for each QoL outcome in Figures 1 to 4. By nature of the methods each utilised a different number of patients and this information is shown on the figures. For example, ANCOVA of 12 month immediate responses for the RQLS used 121 patients, but including the responses received reminder this increased to 276 patients. Imputation or use of a repeated measures approach increased the number of patients used even further. Some differences occurred between QoL scores for the same analysis method due to the problem of missing items within a particular QoL instrument contained within the questionnaire.

Reflux specific QoL (RQLS)

Figure 1 shows the results of the various analysis approaches for the reflux specific QoL score (RQLS). In this situation all the different analysis strategies gave significant estimates of treatment difference, with the surgical procedure providing better follow up QoL scores than those on medical management. The magnitude of this difference did however differ between the analysis strategies, as did the number of participants included in the analysis. Within a particular method, the estimate based on immediate data only was

always lower than that based on all the observed data. This suggests that ignoring the reminder responses, under-estimates the treatment difference and introduces a bias to the results.

To our knowledge there is no published information on what magnitude of change on the RQLS would represent a clinically significant difference. The confidence interval for the smallest treatment difference estimated using LVCF on immediate data included effects of less than 0.2 standard deviations (SDs) of the scores. Using the suggestion by Cohen that 0.2SDs is a small difference, implies that for this study despite statistical significance, we cannot rule out clinically insignificant findings [19, 20]

The number of participants used in the analysis also varies between methods by nature of what they are. The first method (ANCOVA on immediate responders at 12 months) used only 121 (34%) of participants compared to using multiple imputation in addition to all observed responses (99%). The only reason this is not 100% under multiple imputation is because of some missing covariate data. The repeated measures approach (linear mixed effects model) on all responders used 327 (92%) of participants but alongside MI has the assumption of MAR which was shown to be likely [12].

SF12

The results for the SF12 physical summary score are shown in Figure 2 and SF12 mental summary score in Figure 3. The results of the different analysis approaches for the physical summary score follow the same pattern as for the RQLS. All estimates are significant, but of different magnitude and precision, using different numbers of participants. The estimates using the ANCOVA on all data and that obtained under multiple imputation or the linear mixed effects model are similar. Although all statistically significant, the

approach chosen may have an impact on clinical significance. Osoba *et al*., referred to a little change on the SF12 as between 5 and 10 units, with 10-20 as moderate change and clinical significance was regarded as 10 units [21]. The estimates here are all below five so in this instance clinical significance is not affected by the choice of analysis. For the mental summary score (Figure 3), each analysis approach yields a non-significant estimate of treatment difference, but the magnitude differed and the precision varied between them as was seen for the physical summary score.

EuroQoL EQ5D

The EQ5D is the interesting QoL score for this set of participants as the choice of analysis approach did impact on whether a significant difference between treatment groups was found. Figure 4 displays the estimates alongside their 95% CIs. LVCF on immediate data followed by the ANCOVA and a linear mixed effects model on the immediate data both yield statistically significant results ($p = 0.005$ and $p = 0.013$ respectively). The remaining approaches provide non-significant results ($p>0.05$) although the linear mixed effects model on all available data is borderline ($p=0.053$). An estimate of the minimally important clinical difference for the EQ5D has been found to be 0.074 [22] or the slightly higher 0.082 from the more recent paper [23]. In this instance the other method which yields a clinically significant difference between the groups is the linear mixed effects model on immediate data (estimate = 0.084). Using LVCF on immediate data followed by the ANCOVA yielded a statistically significant result but this was not clinically significant. This highlights that the choice of analysis approach can generate a different result based on statistical significance and clinical significance.

Summary of results

Across all four QoL outcomes, the various analysis approaches provided different estimates of treatment difference, with varying precision, using different numbers of patients. The main findings were that the use of the additional reminder data was useful and definitely recommended where possible. Of the different statistical analysis approaches considered and in the presence of missing data at random a linear mixed effect model or multiple imputation were preferred. Use of simple imputation is not recommended. In our opinion the most optimal strategy would be to collect as much information as possible, through the use of reminders (or alternative data collection strategies). Following this a linear mixed effect model or multiple imputation would be suitable.

Discussion

The aim of this paper was to illustrate (using REFLUX) how different choices of analysis methods can impact upon a trial conclusion. Data from three QoL instruments (four outcomes) collected at three time points were obtained. Eight different analysis strategies were implemented and included the original published ANCOVA, a linear mixed effects model, simple imputation (using LVCF) and multiple imputation (using predictive mean match model) followed by ANCOVA. It was found that the choice of method had a bearing on the potential trial conclusion. In this example trial, the conclusion for the statistical significance of the primary outcome (RQLS) would not have been affected (all results remained statistically significant). However, using the approach from Cohen that 0.2SDs can be regarded as a small difference the interpretation of the clinical significance of the difference between the two treatment groups may have been altered [19, 20]. Statistical significance of the SF12 outcomes was not affected by the approach used, but in some cases clinical significance was. For the EQ5D outcome, methods differed in clinical

and statistical significance. Thus across the four QoL outcomes, the work does highlight the fact that you may get a change in conclusion (either statistical significance or clinical significance) depending on the choice of analysis method.

Previous work showed that the missing data in REFLUX was either MCAR or MAR depending on the QoL outcome or time point [12]. Knowing this suggests that the simple imputation methods are likely to provide biased results. Either the linear mixed effects model or multiple imputation process would be more appropriate as they have the assumption of MAR.

It is common practice now in clinical trials to specify the analysis plan in advance, and this type of sensitivity analysis on the trial result should not be undertaken post-hoc. Researchers should pre-specify what they plan to do about any potential missing data, to prevent a subsequent suspicion that they may have tried various methods of imputation and selectively chosen to report the one that gives results most to their liking. This might take the form of pre-specifying a number of analyses to act as a sensitivity analysis to the primary analysis approach.

The choice between different approaches for missing data can also depend on the amount of data missing. Schulz and Grimes give a general rule of thumb with regard to missing data [24]. They suggest that in a trial with less than 5% missing, the bias will be minimal. A trial with over 20% missing poses a serious threat to the validity of the study. In between 5% and 20% missing leads to intermediate levels of problems. This general rule can be applied alongside the approaches set out in this thesis. Imputation is often only regarded as a plausible option when the amount of missing data is less than 20%. Undertaking imputation with more than 20% missing should be done so with caution, as it

is likely that the result of the trial would not be accepted by the research community. This is provided as a guideline and not a rule for all scenarios.

Conclusion

In conclusion, researchers should carefully consider how best to analyse a study where missing data may be an issue. Since the choice of methods may provide different results, the methods chosen should be pre-specified in the trial protocol. Ensuring the maximum amount of data as possible is used is important. Use of reminders to recover data initially missing may be helpful. In addition taking into account all available data (e.g. linear mixed effects model) may be of benefit as everyone with at least one assessment can be included and the assumption of MAR may be plausible. Excluding some people may introduce bias to the results. Imputation is preferred over complete case analysis, as it takes into account all participants within the trial. However, if the proportion of missing data is high, imputation must be used with great caution and the conclusions from the analysis must be regarded as suspect.

Abbreviations

ANCOVA – Analysis of Covariance; BMI – body mass index; CI – confidence interval; EQ5D – EuroQoL EQ5D; GORD- gastro-oesophageal reflux disease; LVCF – last value carried forward; MAR – missing at random; MCAR – missing completely at random; MCMC – Markov Chain Monte Carlo; MI – multiple imputation; MNAR – missing not at random; QoL – quality of life; RCT – randomised control trial; REFLUX - Randomised Evaluation oF Laproscopic sUrgery for reflux; RQLS – Reflux specific quality of life score; SD – standard deviation; SF12 – Short Form 12.

Acknowledgements

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, Samantha Wileman who assisted with data queries and provided background information on the trial.

References

(1) Pocock SJ. Clinical Trials: A Practical Approach. John Wiley & Sons; 1983.

(2) Rubin DB. Inference and missing data. Biometrika 1976; 72:359-364.

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

(4) Grant AM, Wileman SM, Ramsay CR, Mowat NR, Krukowski ZH, Heading RC, et al. Minimal access surgery compared with medical management for chronic gastro-oesophageal reflux disease: UK collaborative randomised trial. BMJ 2009; 337:a2664.

(5) Grant A, Wileman SM, Ramsay C, Bojke L, Epstein D, Sculpher M, 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.

(6) Macran S, Wileman S, Barton G, Russell I, REFLUX trial group. The development of a new measure of quality of life in the management of gastro-oesophageal reflux disease: the Reflux questionnaire. Quality of Life Research 2007 Mar; 16(2):331-343.

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

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

(9) Fielding S, Maclennan G, Cook JA, Ramsay CR. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. Trials 2008 11 Aug 2008;9(51).

(10) Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez N, et al. The European Organisation for research and Treatment of Cancer QLQ-C30: A quality of life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute 1993(85):365-376.

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

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

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

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

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

(18) Fielding S, Fayers P, Ramsay C. Predicting missing quality of life data that were later recovered: an empirical comparison of approaches. Clinical Trials 2010; 7:333-342.

(19) Cohen J. Statistical Power Analysis for the Behavioural Sciences. London: Academic Press; 1969.

(20) Norman GR, Sloan JA, Wyrwich KW. Interpretation of Changes in Health-related Quality of Life: The Remarkable Universality of Half a Standard Deviation. Medical Care 2003;41(5):582-592.

(21) Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of- life scores. Journal of Clinical Oncology 1998 Jan;16(1):139-144.

(22) Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Quality of Life Research 2005; 14:1523-1532.

(23) Luo N, Johnson JA, Coons SJ. Using Instrument-Defined Health State Transitions to Estimate Minimally Important Difference for Four Preference-Based Health-Related Quality of Life Instruments. Medical Care 2010; 48(4):365-371.

(24) Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet 2002;359: 781-785.

Tables

Table 1: Patient characteristics at recruitment

Patient Characteristic	Total (N=357)	Surgical (N=178)	Medical (N=179)
Baseline questionnaire returned – N (%)	349 (98)	175 (98)	174 (97)
Age – mean (SD)	46.3 (11.1)	46.7 (10.3)	45.9 (11.9)
Male – N (%)	236 (66)	116 (65)	120 (67)
BMI – mean (SD)	28.4 (4.2)	28.5 (4.3)	28.4 (4.0)
Duration in months of prescribed medication for GORD - median(IQR)	32 (15,76)	33 (15,83)	31 (16,71)
Employment status - N (%)			
Full-time	226 (63)	116 (65)	110 (61)
Part-time	29 (8)	13 (7)	16 (9)
Retired	34 (10)	12 (7)	22 (12)
Other	68 (19)	37 (21)	31 (17)
Age left full-time education – N (%)			
16 and under	218 (62)	110 (63)	108 (61)
17-19 years	78 (22)	38 (22)	40 (23)
20 years +	58 (16)	28 (16)	30 (22)
Current Smoker – N (%)	86 (24)	46 (26)	40 (22)
Erosive oesophagitis – N (%)	182 (59)	85 (55)	97 (62)
Co-morbidity - H. Pylori status – N (%)			
Positive (subsequently treated)	26 (10)	12 (9)	14 (10)
Negative (subsequently untreated)	4 (2)	1 (1)	3 (2)
Negative	148 (55)	75 (56)	73 (54)
Uncertain	90 (34)	45 (34)	45 (33)
Hiatus Hernia present – N (%)	196 (59)	94 (57)	102 (60)
Asthma – N (%)	42 (12)	21 (12)	21 (12)
Source of recruitment – N (%)			
Retrospective	167 (49)	84 (49)	83 (48)
Prospective	176 (51)	87 (51)	89 (52)

Table 2: Pattern of missing data

Pattern	N (%)	%	Baseline	3 months	12 months
1	290	81.2	-	-	-
2	13	3.6	-	-	x
3	24	6.7	-	x	-
4	4	1.1	x	-	-
5	22	6.2	-	x	x
6	1	0.3	x	-	x
7	1	0.3	x	x	-
8	2	0.6	x	x	x

- questionnaire returned; x questionnaire missing

Figures

Figure 1: Estimates of treatment difference (95% CI) in RQLS at 12 months

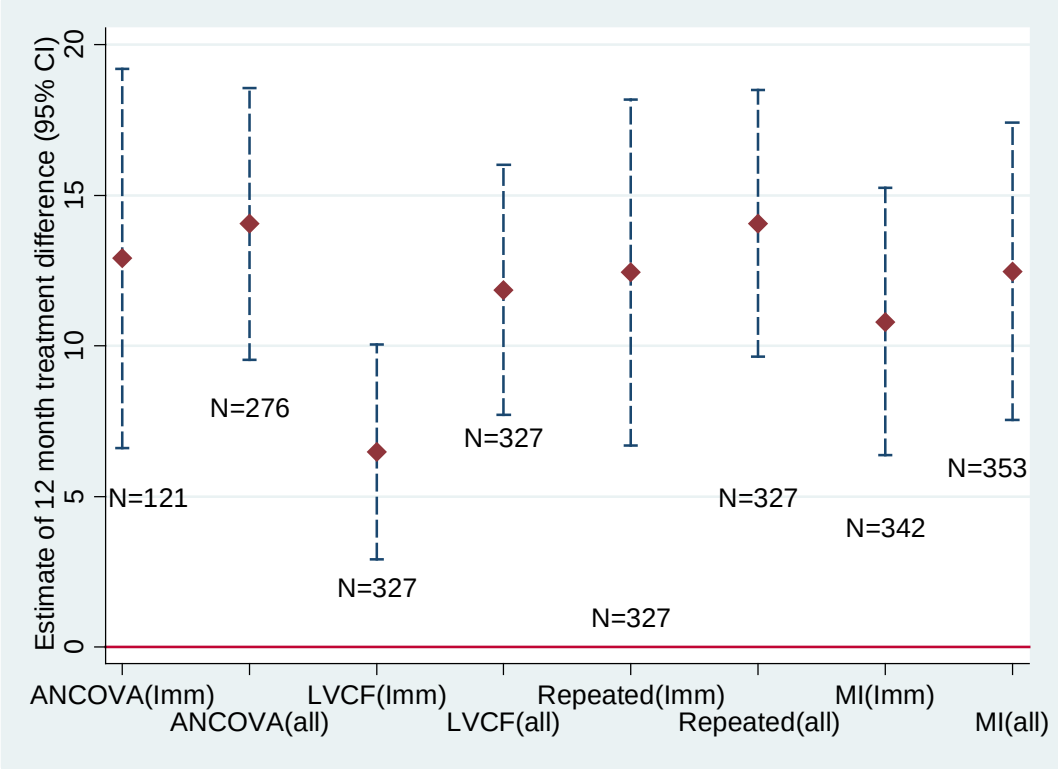

Figure 2: Estimates of treatment difference (95% CI) in SF12 physical summary scores at 12 months

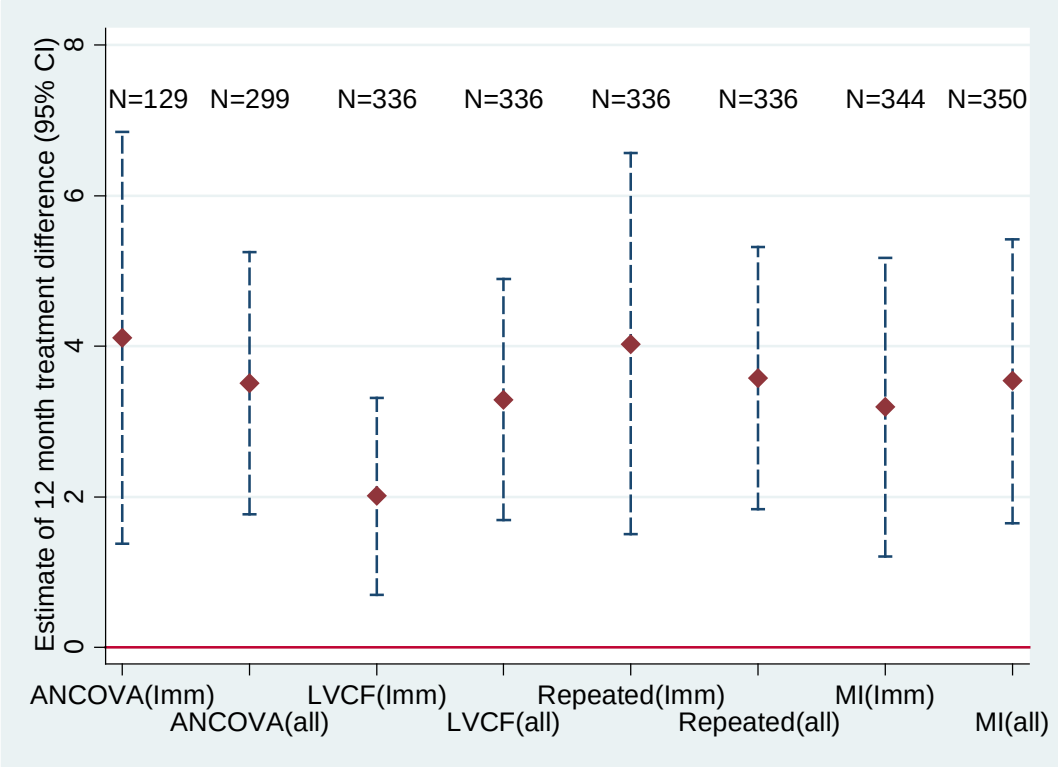

Figure 3: Estimates of treatment difference (95% CI) in SF12 mental summary scores at 12 months

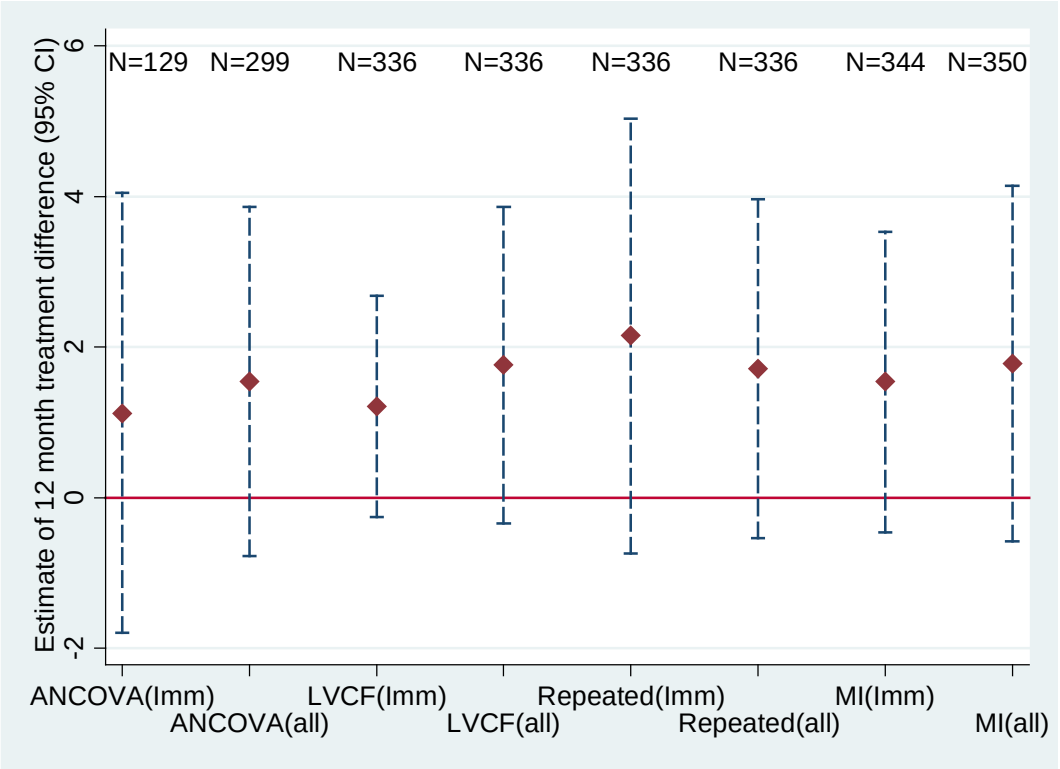

Figure 4: Estimates of treatment difference (95% CI) in EuroQoL EQ5D at 12 months