Practical and statistical issues in missing data for longitudinal patient reported outcomes

Melanie L Bell and Diane L Fairclough

Stat Methods Med Res published online 19 February 2013
DOI: 10.1177/0962280213476378

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–20

DOI: 10.1177/0962280213476378
smm.sagepub.com
SAGE

Practical and statistical issues in missing data for longitudinal patient reported outcomes

Melanie L Bell[1] and Diane L Fairclough[2]

Abstract

Patient reported outcomes are increasingly used in health research, including randomized controlled trials and observational studies. However, the validity of results in longitudinal studies can crucially hinge on the handling of missing data. This paper considers the issues of missing data at each stage of research. Practical strategies for minimizing missingness through careful study design and conduct are given. Statistical approaches that are commonly used, but should be avoided, are discussed, including how these methods can yield biased and misleading results. Methods that are valid for data which are missing at random are outlined, including maximum likelihood methods, multiple imputation and extensions to generalized estimating equations: weighted generalized estimating equations, generalized estimating equations with multiple imputation, and doubly robust generalized estimating equations. Finally, we discuss the importance of sensitivity analyses, including the role of missing not at random models, such as pattern mixture, selection, and shared parameter models. We demonstrate many of these concepts with data from a randomized controlled clinical trial on renal cancer patients, and show that the results are dependent on missingness assumptions and the statistical approach.

Keywords

Missing data, maximum likelihood estimation, generalized estimating equations, multiple imputation, quality of life, patient reported outcomes, cancer

1 Introduction

Patient reported outcomes (PROs) are used regularly in clinical trials and other types of health research. They have played a significant part in the development and evaluation of preventive, psycho-social, and therapeutic interventions.[1] In clinical trials, PROs have been used for choosing the best treatment, e.g., where two treatment regimes may be expected to give similar survival outcomes but have differing side-effect profiles; for understanding the patient experience, e.g., for offering counselling; and for improving clinical trials, e.g. when PROs are used in prognostic modelling.[2]

[1]The Psycho-Oncology Co-operative Research Group (PoCoG), University of Sydney, Sydney, Australia
[2]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, Colorado, USA

Corresponding author:
Melanie L Bell, School of Psychology, Transient Building (F12), University of Sydney 2006, NSW, Australia.
Email: melanie.bell@sydney.edu.au

PROs are used to subjectively assess outcomes that are difficult or impossible to measure physically, such as satisfaction with care, pain, depression, or quality of life. They are also used for symptoms of disease or treatment toxicity, like vomiting, sleeplessness, or dyspnea, which could be physically measured, but with more difficulty and/or measurement error. PROs are typically measured with questionnaires, which are made up of *items* (questions) grouped into one or more *subscales*, or *domains*, which purport to measure some underlying construct. For example, a widely used PRO for quality of life (QoL) assessment in cancer is the FACT-G, which is made up of 4 subscales: physical, emotional, functional, and social[3] and disease or symptom-specific subscales added as needed. Desirable characteristics of PROs are validity, reliability, and responsiveness to change. Most PROs in use have undergone psychometric analysis to assess these qualities and are as reliable (if not more) than many physically measured outcomes.[4]

Many studies which use PROs are longitudinal and are therefore likely to have missing data. For PROs this can be individual questions, termed *missing items*, or missing questionnaires, as might occur if a patient did not attend an assessment. There are two important potential consequences of missing data. The first is the decrease in precision (wider confidence intervals) and power caused by the reduction in data. The second, and more serious, is the potential for bias in the estimation of both between (e.g., treatment effect) and within group effects (e.g., change over time). In therapeutic oncology trials, for example, patients who are sicker may be less likely to complete QoL questionnaires. If this occurs, QoL will be overestimated and toxicity underestimated. In a trial assessing a decision aid for genetic testing, those with higher decisional conflict may be less likely to return follow-up questionnaires designed to measure decisional conflict and regret. If this is the case, and the control arm (no decision aid) has higher decisional conflict than the arm which received the decision aid, the estimate of treatment effect will be attenuated.

Robust methods for missing data have been an active area of research in the statistics literature for the last few decades. This has not translated well into the application of statistics, however. Inappropriate methods for handling missing PRO data in longitudinal randomised controlled trials (RCTs) have been shown to be the norm in the top medical journals[5,6] as well as in specialist fields, such as psychology and psychiatry[7–10] and epidemiology.[11,12] Recent surveys of journal editors show that improper methods for missing data is a top concern.[13–15] Reporting on missing data is poor,[6] with one study estimating that 65% of studies indexed in PubMed do not report how missing data are handled.[16] Assessing the sensitivity of results to assumptions about missing data (sensitivity analyses) are rarely performed, and when they are, they are usually inadequate.[6]

Although missing data has the potential to cause serious bias, it is still possible to perform a valid and sensible analysis.[17–24] In order to do so, however, attention must be given at each stage of the research: design, data collection, statistical analysis, and reporting. The aim of this paper is to address the issue of missing data in PROs, including key definitions, prevention practices, and analytical approaches, including sensitivity analyses. We will not discuss questionnaire development, and make the implicit assumption that the PRO used is psychometrically reliable and valid in the target population.

2 Example

We demonstrate concepts and methods with data from a multicenter randomized phase III trial comparing two treatments in advanced renal-cell carcinoma patients, which had high rates of missing data. The treatments were 13-cis-retinoic acid plus interferon alpha 2a (experimental) and interferon alone (control).[25] There were 284 patients that were randomized to the experimental treatment or the comparison treatment; 230 patients were invited to join the QoL sub-study and

Table 1. Number of observations at each time-point for the renal cancer trial.

	Control arm				Experimental arm			
Weeks	0	0–4	4–13	13–26	0	0–4	4–13	13–26
Time	1	2	3	4	1	2	3	4
# obs	97	92	70	35	100	86	61	30

213 returned at least one questionnaire. The dataset used here is a (stratified by treatment) random sample of 197 patients. Patients were assessed at 6 time-points: baseline, 2, 8, 17, 34, and 52 weeks. By the fourth assessment (17 weeks), only 43% of surviving patients had complete QoL data, which was 35% of all patients, see Table 1. By 34 and 52 weeks there was very little data, and we do not use these values.

The trial used the FACT-G to assess general QoL as well as disease-specific symptoms. We consider the Trial Outcome Index: a sum of the physical and functional well-being scores and 17 disease-specific items. It has been scaled to a range of 0–100, with higher values indicating better QoL. The minimum important difference for the FACT-G has been estimated to be between 5 and 7.[26] Additional data that were collected were survival (yes/no) and survival time; progression (yes/no) and time to progression (progression refers to worsening cancer); time on treatment; and baseline risk score which was based on physical measurements, and had values of low, medium, and high.

To focus the discussion, we estimate the mean difference in QoL at the fourth time-point ($QoL_{exp} - QoL_{ctl}$) using various approaches. Treatment estimates are summarised in Table 4, and within-group time trajectories are shown in Figure 2. All mixed models used the Kenward-Roger degrees of freedom, which performs well for both small and large samples.[27] All generalized estimating equations (GEEs) assumed an independent working correlation matrix. To make comparisons between methods easier, we imputed the data with a single EM imputation[28] to make the incomplete data monotone (about 1% of the entire dataset). All analyses were performed in SAS v9.2.

3 Missing data definitions and descriptions

3.1 Missing items

One source of missing data is missing responses to individual questions. Many questionnaires have instructions regarding scoring when some items are missing, and the tool may not be valid or reliable unless these are followed. A common approach is half-mean imputation: if half or more of the items are complete, the missing items can be imputed with the mean of the remaining items. If there is more than one subscale, this is done within each subscale. In general, this is a valid approach because psychometrically validated questionnaires measure some underlying construct, so items within the questionnaire are correlated with one another. Fairclough and Cella[29] investigated various approaches to missing items and found that this fairly simple approach yielded robust results. However, if there is an ordering of difficulty, this may not be appropriate.[19,30] For example, in a physical functioning scale, if patients have difficulty getting out of bed, then they certainly will have difficulty climbing stairs. In these cases more complex rules for imputation may be needed. In general, missing items are infrequent and missing subscale scores due to missing items are even less frequent. In the event of a high proportion of missing items, the appropriateness of the particular questionnaire in the population being studied should be examined.

3.2 The missingness hierarchy

Missing questionnaires are more common and potentially have an impact on analysis and interpretation. Rubin defined a taxonomy of three types of missingness.[31] Data missing completely at random (MCAR) are those where the probability of missing is unrelated to the patient's outcome. For example, a researcher forgets to administer the questionnaire. Data missing at random (MAR) are those whose probability of being observed depends only on past observed data (outcomes and possibly covariates). For example, a doctor tells the patient not to complete the questionnaire because he was too sick at the last visit. The essential concept of MAR is that *conditional on observed data*, the data are MCAR. Data missing not at random (MNAR) are those whose probability of missingness depends on the value of the missing outcome itself, *even when observed data are taken into account* (both outcomes and covariates). Sometimes a distinction is made between MCAR and covariate dependent MAR. However, when the covariate is conditioned on (by including it in the model) data become MCAR, so we do not distinguish between these types. The *missingness mechanism* is the underlying cause of why data are missing. It has been described as a "second stage of sampling", so MCAR data would be the result of simple random sampling.[17] For MAR and MNAR data, the sampling scheme is unknown, which is why selection bias can occur. The term *ignorability* roughly means that the data are MCAR or MAR. It does not mean that one can ignore missing data; it means that the data can be analysed without explicitly modelling the missingness.

3.3 Patterns of missing data

If a patient drops out of the study or dies, their data from a certain point onwards will be unobserved. This pattern of missingness is termed *monotone*. *Intermittent* missingness is defined to be when an outcome is unobserved at one assessment but is observed at a following assessment. Intermittent missing data are more typical in studies of individuals with chronic conditions. In contrast, monotone missing data occur in studies of populations with significant morbidity or mortality, and in trials where monotonicity occurs as the result of the design.

3.4 Dropout and death

The question of how to deal with patients who die is a "vexing issue" which has not been resolved.[32] Some researchers impute a score of 0 (or whatever is the minimum possible score on the scale). While this is reasonable for some scales where 0 is explicitly anchored to death (e.g., utilities, functional well-being), it does not make sense for others such as symptom and physical scales, where a 0 could mean that the deceased patient is experiencing severe nausea, vomiting, and pain. For a perspective on causal inference in the presence of death see Ref. 33. For an overview of different approaches see Ref. 34. It is important that researchers are aware that there are either conceptually or analytically an implicit or explicit imputation that is made for every missing observation. In the renal example, we do not distinguish between monotone missing data due to dropout versus death.

3.5 Mathematical definitions and notation

Suppose there are N independent patients with n planned assessments, and $n_i \leq n$ observed values for patient i. Let $R_{ij} = 1$ if the outcome, Y_{ij}, is observed for patient i at time j. Let X represent

covariates (such as treatment assignment). The vector of outcome observations can be partitioned into $Y = (Y^O, Y^M)$, where Y^O is the observed data and Y^M is the missing data. Then

$$\begin{aligned}
\text{MCAR} = &> P(R_{ij}|Y_{ij}, Y_{ij-1} \ldots Y_{i1}, X) = P(R_{ij}|X) = P(R|X) \\
\text{MAR} = &> P(R_{ij}|Y_{ij}, Y_{ij-1} \ldots Y_{i1}, X) = P(R_{ij}|Y_{ij-1} \ldots Y_{i1}, X) = P(R|Y^O, X) \\
\text{MNAR} = &> P(R_{ij}|Y_{ij}, Y_{ij-1} \ldots Y_{i1}, X) = P(R_{ij}|Y_{ij}, Y_{ij-1} \ldots Y_{i1}, X) = P(R|Y^M, Y^O, X).
\end{aligned}$$

The likelihood of the complete data (without loss of generality, excluding X) is

$$f(Y^O, Y^M, R) = f(Y^O, Y^M) f(R|Y^O, Y^M),$$

and of the observed data is

$$f(Y^O, R) = \int f(Y^O, Y^M) f(R|Y^O, Y^M) dY^M.$$

The integral averages over all possible values of Y^M, weighted by the probability of their occurrence. If data are MAR, then

$$\begin{aligned}
&f(R|Y^O, Y^M) = f(R|Y^O), \text{ so} \\
&\quad f(Y^O, R) = f(R|Y^O) \int f(Y^O, Y^M) dY^M = f(R|Y^O) f(Y^O).
\end{aligned}$$

$f(R|Y^O)$ gives no information about Y^0, so it can be ignored in the analysis. I.e., if data are MAR, the missingness mechanism does not need to be included in the modelling. A similar argument can be made for MCAR data.

4 The design stage: minimizing missing data and collecting auxiliary data

While some missing data in longitudinal studies is nearly inevitable, there are ways to minimize it.[2,35,36] The first step is literally the first step: in the design. Whether or not the PROs are a primary or secondary outcome, they need to be fully integrated into the design and the conduct of the study, with this integration codified in the study protocol, quality assurance measures, and statistical analysis plan. One design decision that impacts on missing data is the decision to continue assessments after the patient misses an assessment or goes off treatment. In some settings, the impact of treatment on the outcome will occur after treatment failure. Continued assessment is conservative; if it is decided later that these data are not relevant to the research question they can be excluded. This recommendation is balanced with a caution about the length of follow-up in populations with high rates of morbidity. Assessments should generally not be planned after the median survival (and possibly should be a shorter interval).

Missing data and sensitivity analyses (described below) should also be addressed in the protocol's statistical analysis plan. This ensures that adequate resources, in terms of time, data management, and statistical programming will be allocated so that robust results are more likely. Quality assurance and quality control in the setting of longitudinal studies with PROs should focus largely on minimizing missing data. If the study has multiple centres, assign a PRO coordinator who is in charge of training, monitoring compliance rates, and PRO quality assurance. The protocol should contain justification of the PROs, including how they address the objectives of the trial. This fosters buy-in by participating centres, which is essential to reducing missing data rates. Build in reminders for participants to complete their assessments; consider phone calls, text messages, and

email. If possible, reduce participant burden, by using shorter questionnaires and helping with travel and childcare.

4.1 Auxiliary and covariate data

Most studies collect additional data beyond the primary outcome. These covariates may be of direct interest, or may be collected to control for confounding, to improve precision in the estimates, for modelling interactions, or for handling missing outcome data. Data that are used in this last way are often referred to as auxiliary data.[37,38]

Incorporation of auxiliary data can make assumptions about the ignorability of the missing data more likely. Auxiliary data is useful if it is predictive of missingness, of the outcome, or both.[17,19,38,39] It is possible that many trial statisticians are not aware of the possible benefits of using auxiliary data to improve estimates in the face of missing data, despite the recognition of these in the statistical literature.[40] In a review of covariate adjustment in the top medical journals, none of the reasons given for covariate adjustment had to do with missing data.[41] Some regulatory agencies also do not mention the use of auxiliary data for missing data when they discuss covariate adjustment in clinical trials[42] nor do two popular texts on clinical trials.[43,44]

Two common ways of using auxiliary data are to (1) include them in the model either as a covariate[17,45] or (2) use them in a multiple imputation (MI) model. Variables measured post-baseline should generally not be used as covariates in the model, because they can attenuate the treatment effect (by adjusting it away) and make estimates difficult to interpret. However, post-baseline auxiliary covariates can be used in MI. Two other approaches that have been used in the statistics literature are (1) for creating weights for weighted GEE[46] or stratified propensity (to be missing) score analysis[40] and (2) joint modelling of the auxiliary and outcome data.[17,38] Auxiliary data can also be used as an important part of sensitivity analyses (discussed below), as a vehicle for exploring the missingness mechanism, and arguing for the missingness assumption. These uses of auxiliary data are discussed later in more detail.

In anticipation of unavoidable missing data, auxiliary data should be collected, such as proxy measures of the PRO of interest and/or variables that may correlate with reasons for dropout and the PRO. For example, if collecting QoL in a clinical trial, also collect Eastern Cooperative Oncology Group(ECOG) performance status which is a scale filled out by a clinician to monitor physical function which ranges from 0 (perfect health) to 5 (death). Require, as part of the trial quality assurance, that a form such as Table 2 be filled out for every planned assessment.[19] Careful wording of the response will allow the analyst to indicate which are likely to be related to the outcome.

The renal trial was designed prior to these recommendations. As previously mentioned, it did include a baseline measure of risk, as well as post-baseline assessments of time on treatment, time to disease progression, and death. In retrospect, it would have been advisable to collect a longitudinal measure such as the ECOG performance status.

5 First steps: Describing the extent and patterns of missingness

5.1 Simple descriptions of missing data

After the study has been conducted and it is time to report results, the first step is to describe how many participants were in the study at each time-point. A table, such as Table 1, or a CONSORT flow diagram[47] is a good way to achieve this. In the renal cancer trial, the amount of intermittent

Table 2. Items to include on a form monitoring completion and reasons for missing outcome data.

Questionnaire complete: Yes_____ No_____
Main reason for non completion 1 = Patient felt too ill 2 = Staff felt patient too ill 3 = Patient refused for non-health reasons 4 = Unable to contact 5 = Administrative oversight 6 = Other (please specify) 7 = Unknown

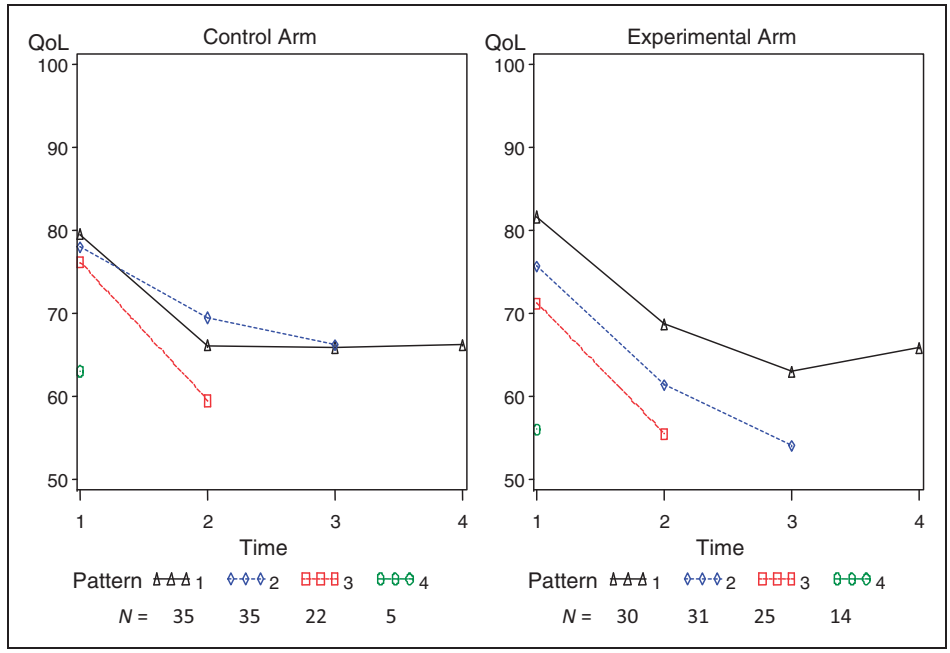

Figure 1. Missingness patterns: quality of life stratified by treatment group and dropout time. The possible range of QoL is 0–100, with higher values indicating better QoL.
QoL: quality of life.

missing data was 1.1% of the total number of assessments which we used (197×4), so a more detailed description is not presented.

To explore mechanisms of missingness, a good place to start is a graph of the PRO versus time, stratified by dropout time, as shown in Figure 1 for the renal data. If the trajectories over time are substantially different, data are not MCAR. For example, are patients who have lower baseline values more likely to drop out, or are steeper rates of increase or decrease over time associated with dropout? The separation between the missingness patterns is more dramatic in the experimental arm than control, thus we would expect more sensitivity across analyses in this arm.

Table 3. Correlation[a] of dropout with other covariates, and results of *t*-tests comparing those who drop out (n = 132) versus those who do not (n = 65).

Variable	Mean difference (no dropout – dropout)	95% CI	*p* Value	Correlation with dropout
Baseline QoL	11.2	(7.6, 14.9)	<.0001	−0.19
Risk group (low, med, high)	−0.30	(−0.51, −0.10)	0.025	0.16
Time on treatment (days)	150.9	(101.0, 200.8)	<.0001	−0.46

[a]Kendall's tau.

5.2 Missing data models to explore predictors of missingness

Exploration of the missingness mechanism(s) can be performed comparing those who drop out versus those who do not, via t-tests, cross tabulations, or logistic regression. Survival analysis can be used to look at predictors of time to dropout. The decision to embark on an analysis should be driven by its goals; whether the intent is to describe the characteristics associated with missing data or to identify alternative analytic methods. Demographic and baseline characteristics are often weakly associated with missing data, and while these associations may be statistically significant, they typically explain less than 10% of the variation (e.g. $R^2 < 0.1$). This combined with similarly weak correlations with the PROs results in minimal usefulness in any analytic strategy. In contrast, proxy measures of the outcome and other clinical outcome may be strongly correlated with both missingness and the outcome and thus may be useful as auxiliary variables in imputation and joint models (more details later).

Table 3 compares those who dropped out (at any time) with those whose data were complete for the renal trial. There were statistically and clinically important differences for all variables, with those who dropped out having worse baseline QoL, shorter survival time, treatment time, and a higher risk group on average. Figure 1 shows missingness patterns that contrast the scores across the different times of dropout; the difference between the curves confirming that the data are not MCAR.

6 Analysis approaches to avoid

As mentioned in the introduction, most RCTs have missing data, and nearly none use modern missing data methods. Many statistical techniques used by researchers in various fields make the strong assumption that their missing data are MCAR,[5,7,10,13–15] but this is unlikely for many situations. In the following section, we detail which analyses make this assumption as well as discussing other approaches to avoid.

6.1 Simple imputation and LOCF

Simple imputation refers to filling in missing observations with a single value. Common choices are the mean of the sample (marginal mean), the individual's last observation carried forward (LOCF), their mean, best or worst values, or an extrapolated or interpolated value based on a regression on the individual (conditional mean). There are two problems with simple imputation. First, variance in

the sample is artificially reduced, leading to underestimated standard errors and increased type I error rates. Second, simple imputation can yield biased results. LOCF, for example, can be biased in unpredictable ways, because it makes unrealistic assumptions about individuals' time trajectories.[5,9,17,19,22,23] For QoL measures, under LOCF, patients who drop out would get the same QoL scores as their last assessment's QoL, which is improbable in many situations (like advanced cancer). Although simple imputation has been advocated as a small part of a sensitivity analysis,[19] others have argued that even in this context it should not be used.[17] Molnar gives evidence that the use of LOCF may have resulted in more toxic dementia therapies being favoured over less toxic therapies.[9]

6.2 Analyses which assume data are MCAR

6.2.1 Complete case analysis

A complete case analysis occurs when only those participants with complete data for all assessments are analysed. The widely used multivariate analysis of variance (MANOVA) and repeated measures ANOVA (depending on software) are both forms of a complete case analysis, and their use has been discouraged by various statisticians.[10,23] Estimates will be unbiased only if data are MCAR, but even then, a lot of data is thrown away, which is unethical and statistically inefficient. For example, in oncology trials it will generally overestimate QoL and underestimate toxicity over time. It makes strong assumptions about the covariance structure of the data (that it is compound symmetric), an assumption that is relaxed in the mixed model repeated measures (MMRM) analysis described later.

The selection bias is illustrated for the renal-cell carcinoma trial in Figure 2, which shows how the QoL trajectory over time is overestimated for the experimental arm for the complete cases (only 65 of the 197 patients). The difference in QoL between the experimental and control groups at the fourth time-point is small: -0.36 (two sample t-test, $p = 0.9$) or -2.22 (ANCOVA, conditioning on baseline QoL, $p = 0.4$). The small differences between the two arms are not unexpected as we are comparing only the *selected* patients who remain on the trial until the fourth assessment.

6.2.2 Repeated univariate analyses

Repeatedly testing at each time-point is common but suboptimal. In addition to producing biased results (if data are not MCAR) by ignoring observed data at other points in time, it also does not take advantage of the longitudinal nature of the data, by comparing different groups of patients at each time-point.

6.2.3 Unweighted generalised estimating equations

GEE is a frequentist method which can be used to analyse normal and non-normal longitudinal data.[48–50] In GEE, the mean and covariance are modelled separately, which leads to the appealing feature that even if the covariance is incorrectly specified, the parameter estimates involved in the mean are still unbiased asymptotically, as long as the mean structure is correct. If data are not MCAR, results can be biased.[51] However, in this case weighted methods are often recommended (see below).[46,52]

In our example, estimates from the unweighted repeated measures (RM) GEE were higher than estimates obtained from most of the other approaches (Figure 2). Estimates of the difference at the fourth time-point for the renal data yielded an estimate of -0.36, $p = 0.9$, or -1.54, $p = 0.6$ when time was modelled continuously, using a quadratic (in time) model. (The model included terms for treatment, weeks, weeks^2, treatment $\times$ weeks, treatment $\times$ weeks^2.)

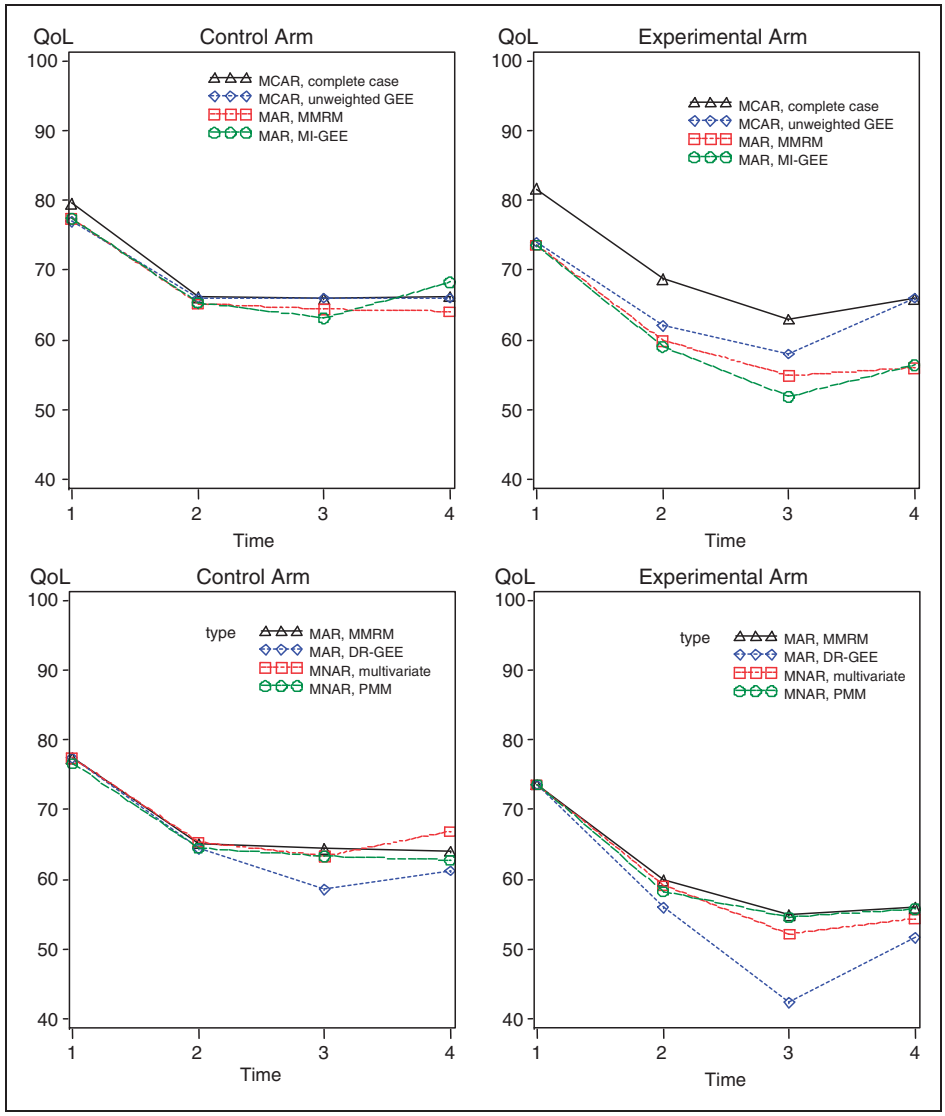

Figure 2. Estimated QoL trajectories within treatment group, by selected analytical approaches. MCAR methods will yield biased estimates both within and between treatment groups.
MMRM: mixed effect regression model; MCAR: missing completely at random; DR-GEE: doubly robust GEE; PMM: pattern mixture model; multivariate: joint model of log time on treatment and QoL; QoL: quality of life.

6.2.4 Summary measures

Summary measures, or derived variables from the raw data, are an approach to simplifying longitudinal data which has been used in various fields.[19,53,54] A set of repeated measures on an individual are reduced to a single value, such as the maximum, the slope, or the area under the time curve. Groups can then be compared using a two-sample t-test or similar. While the simplicity is appealing, an obvious problem is how to determine the value for an individual when some of their

data are missing. Bell et al.[55] have shown that various summary measures approaches can yield biased estimates of intervention effects for all types of missing data.

7 Analyses which assume MAR

7.1 Maximum likelihood methods

Maximum likelihood refers to a parameter estimation method which is used in various longitudinal models including mixed models, latent variable modelling, item response theory and structural equation models (where it is often referred to as full information maximum likelihood). Estimates obtained through maximum likelihood are asymptotically unbiased if data are MAR and the model has been specified correctly.[31]

Mixed models are commonly used for longitudinal data because they allow for non-independent observations and a variable number of observations per patient. Time can be continuous or categorical. When time is included categorically, the model is sometimes referred to as MMRM, means model, response profile analysis, or saturated model. When time is continuous, these models have been referred to as growth curve (GC) models, linear mixed models, and mixed effect regression models. Means models do not assume any particular form for the outcome's time trajectory. Linear mixed models assume a relationship (generally a linear, polynomial, or spline function). Two excellent and readable texts are Refs 23 and 56.

The MMRM is a special case of a means model which models all the covariance within subject using an unstructured covariance.[57] It is not actually a mixed model, as there are no random effects, but is typically referred to as a mixed model because the same software procedures are used for this and true mixed effects models. The generality of this model, both in the mean and covariance structure, protects against misspecification to all but the missing data mechanism and because of this has been recommended for use in the regulatory environment.[57] For all but very small sample sizes, the reduction in power due to use of an unstructured covariance is negligible.[17]

Mixed models, if specified correctly, are valid for data which are MAR. Information from the observed data is used to implicitly impute unobserved data. This implicit imputation is most obvious for the EM algorithm but occurs in all the algorithms that maximize the likelihood. The underlying (and untestable) assumption, therefore, is that the distribution of the unobserved data is well approximated by the distribution of the observed data. Because the imputation is implicit, it is straightforward to implement in standard statistical packages. Note that this implicit imputation relies on exactly the same assumptions as most MI procedures (see below) and results will be very similar.

In contrast to the methods described in the MCAR section, estimates based on mixed models showed a greater decline over time (Figure 2) and larger treatment group difference at the fourth time-point. We fit an MMRM using terms for treatment, time, and the interaction of treatment with time. Baseline QoL was included as one of the repeated outcome measures, thus the 19 patients with only a baseline assessment are included in the analysis. The estimate of the difference in QoL at the fourth time-point between treatment groups for the renal cancer study was -8.08, $p = 0.03$, indicating lower QoL for the experimental group. A quadratic growth curve model was also fit, using fixed effects of treatment, weeks, weeks^2, weeks $\times$ treatment, weeks^2 $\times$ treatment, and random intercept, weeks, and weeks^2. The estimated treatment difference at week 17 was -7.04, $p = 0.008$.

Summary statistics (parameter estimate summary) are related to summary measures, as described in section 6. Instead of reducing the raw data of each subject to a single measure, it reduces the multiple parameter estimates from a model into a single estimate (mean slope, AUC, etc.). The difference is that these summarise group values, not individuals'. An important feature is that there

is no need to specify complex decision rules regarding missing data, as is required for individuals' summary measures, because the implicit imputation of the model takes care of it. Summary measures are only valid for data that are MCAR; summary statistics based on ML methods are valid for data that are MAR.[55] As we are focusing on a single time-point, we do not show either of these for the renal data.

7.2 Multiple imputation

MI is based on filling in missing data by drawing from a distribution of likely values, and does not suffer from the variance underestimation of simple imputation.[58] MI can be a useful way to incorporate auxiliary data when one does not want to use an adjusted model, and can be used for either outcomes or covariates (particularly in observational studies when important covariates are missing). This is achieved by using auxiliary data as covariates in the imputation model (see below).

Suppose there are two observations on each participant: Y_1 (which is completely observed) and Y_2 (which is incomplete). The association between Y_1 and Y_2 from those participants who have complete data is used to fill in the missing values of Y_2. Mathematically, we draw values of Y_2 from the conditional distribution $Y_2|\ Y_1$. The implicit and un-testable assumption is that the relationship between Y_1 and Y_2 is the same for those who complete and those who do not. This is the MAR assumption.

There are three steps to MI: (1) Impute multiple (M) times, using a regression model called the imputation model, so that there are M complete sets of data; (2) Analyse each of the M data sets; (3) Combine the results using Rubin's rules.[59] These rules take into account the within and between imputation variability, so that the uncertainty associated with imputation is accounted for. The final estimate (regression coefficient, mean difference, etc.) is the mean of the M estimates. The variance is $W + (1 + 1/M)B$, where W is the within-imputation variance and B is the between-imputation variance. These steps are implemented in many statistical packages. For an excellent review of software see Horton and Kleinman.[60] Most multiple imputation assumes that the data come from a multivariate normal distribution, however, the procedures are robust to moderate deviation from normality in typically sized trials ($N \geq 200$).

7.2.1 *Multiple imputation of outcomes versus direct modelling*

If only outcomes are imputed, and the imputation model is similar to the longitudinal model, MI estimates will be very similar to maximum likelihood estimates, and in this case there is little point in performing an MI. However, MI is one strategy for incorporating auxiliary data, particularly when auxiliary data includes post-baseline measures since inclusion as covariates directly into the analysis model will often bias estimates of the relationship between baseline predictors (like treatment) and the outcome toward the null. Also, some researchers prefer to see a simple unadjusted analysis.

7.2.2 *Practical issues with multiple imputation*

While a full review of MI is beyond the scope of this paper, we mention a few key points. First, the longitudinal structure of the data must be accommodated (within subject correlation must be maintained). This can be accomplished by imputing in "wide form" where there is one observation per subject, as opposed to "long form", where there is one observation for each of the time-points of each subject. However, when variables are highly correlated, numerical issues may

arise, which multiple imputation with chained equations (also known as fully conditional MI) may resolve.[60,61] Second, for RCTs the imputation model should include a treatment indicator variable or, even better, be done by treatment group. (The latter enables the analyst's model to examine interactions (effect modifiers) with treatment.) Omission of variables from the imputation model will yield results that are biased towards the null.[19] Third, the analyst's model should not be richer than the imputer's model. I.e., if covariates X_1 and X_2 are used in the analysis, one should not use just X_1 for imputation.[37] If the imputer's model is the same as the analyst's, using multiple imputation for the missing outcome variable will yield similar point estimates as using maximum likelihood estimation, but with slightly larger standard errors for MI.[37,62] An inclusive strategy should be adopted over a more restrictive one.[37]

We illustrate the inclusion of auxiliary data in an MI approach by using Markov Chain Monte Carlo (MCMC) to multiply impute QoL, by treatment, using log survival time, log time on treatment, and risk group. While it is not, in general, advisable to use censored variables in imputation, all but 1 of the 46 censored patients survived past 6 months. Inclusion of other variables (time to disease progression, progression, death) resulted in non-convergence. Although a common recommendation is that only a few (3–5) imputations are needed, in this particular example, with its large proportion of missing data at 6 months, results with 20 imputations were unstable, so we used 100 instead.

An MMRM was then fit to the multiply imputed data, and the treatment estimate at the fourth time-point for the renal cancer RCT was -11.87, $p = 0.02$. This is slightly larger in magnitude to the MMRM without multiple imputation (–8.08), which suggests that the auxiliary variables are contributing additional information. This does not demonstrate that the data are actually MAR.

7.3 GEE extensions: MI-GEE, weighted GEEs, doubly robust GEEs

While estimates obtained by unweighted GEEs are biased for data which are not MCAR, it is still possible to use GEE either with weights, MI, or both: a procedure called doubly robust estimation.[52,63,64] Modelling with weighted GEEs (WGEE) is a two-step process. The first step is to model the probability of being observed to obtain predicted probabilities for each patient. The second step is to fit a GEE, using the inverse of these probabilities as weights. Only observed data are used but are weighted to account for those who drop out. Estimates from weighted GEEs are unbiased for data which are MAR, provided the weight model and the longitudinal outcome model are specified correctly. It is necessary to assume an independent "working covariance" in these models to ensure that the weights are correctly incorporated.[23]

Doubly robust GEEs (DR-GEE) combine inverse probability weighting with MI. They require correct specification of the weight model or the imputation model, but not necessarily both, and will yield unbiased estimation for MAR as long as there are no unmeasured confounders.[52,64,65]

We fit several GEE extension models. The first, an unweighted RM GEE using the multiply imputed data, yielded a treatment estimate of -11.87, $p = 0.02$; nearly identical to the MI-MMRM estimate. We then modelled the cumulative probability of dropout, using the previous QoL, treatment, survival time, days on therapy, baseline risk group, and indicator of death. The inverse probability of being observed at each time-point, for each patient was used as the weight, with weight at baseline $= 1$. The weighted GEE (WGEE) estimate for treatment difference was -6.14, $p = 0.3$. Doubly robust GEE using these weights and MI covariates gave -9.77, but with an SE $= 14.01$, which is over three times that of most other estimates.

7.3.1 Sensitivity of GEE results

There was no evidence of lack of fit for the dropout model, according to the Hosmer and Lemeshow deciles of risk statistic.[66] However, some estimated weights were very large, which can occur when estimated probabilities are small. In the WGEE these large weights, in general, were not used in estimation, as the patients with extreme weights also had missing QoL. For the DR-GEE, however, these weights do get used, as all missing values have been imputed. We examined two strategies to address high weights: replacing all weights > 25 with 25 (the highest 1%) and all weights > 5 with 5 (the highest 5%). The results for the DR-GEE were very sensitive to the strategy adopted. The estimates for the 1% and 5% truncation yielded estimates of -15.38 (with a smaller, although still comparatively large SE $= 8.8$) and -11.08. A pre-specified approach to high weights could prove difficult.

To explore the sensitivity of the results to both the dropout and imputation models, we varied the covariates and fit several combinations. Treatment estimates ranged from -5.7 to -0.6 for WGEE and from -15.4 to 7.3 for the DR-GEE. This, together with the extreme sensitivity to a few observations and large SEs, indicates that weighted GEEs, doubly robust or not, may not be appropriate for these data. Simulation studies have shown that MI-GEEs are fairly robust to misspecification of the imputation model—more so than WGEEs are to misspecification of the dropout model.[52] For a discussion of the limitations of doubly robust GEEs, see Ref. 67.

8 Analyses when data may be MNAR

8.1 Sensitivity analyses

All models for missing data, whether MCAR, MAR, or MNAR, rest upon strong assumptions and can lead the researcher astray if their assumptions or model(s) are incorrect. Thus it is important to try different, clinically plausible models that use different assumptions and see how the estimates change. This is referred to as sensitivity analyses. A thorough and sensible sensitivity analysis is an important step in producing and reporting robust estimates. The National Research Council's Panel on Handling Missing Data in Clinical Trials recently made this recommendation:

Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting.[68]

Ideally, sensitivity analyses should be pre-defined, and written into the protocol and grant proposal. A predefined approach strengthens the scientific rigour and lends credibility to results. In practice, this may be difficult, as some models may not fit (see below). Sensitivity analyses should be based on clinically plausible departures from MAR (e.g., not LOCF). This may best be achieved by discussing this ahead of time with collaborators, and by collecting auxiliary data. They should be understandable to clinical researchers and funders, and be reported along with the primary analysis results.[17] MNAR models must form a key part of a sensitivity analysis if there is any possibility that data are MNAR.

8.2 MNAR models

The major challenge of MNAR models is that for the same reason that we cannot test the MNAR assumption, we cannot demonstrate that any given MNAR model is correct (unless strong assumptions are made). This is primarily because the lack of fit of an MNAR model does not indicate that the data are MAR, rather it means that the data do not fit that particular MNAR

model; it may fit another one. This is why the role of MNAR models are primarily as a part of sensitivity analysis.[22]

There are three main approaches for MNAR models: pattern mixture, selection, and shared random effects models. We review each briefly. Each convert the problem to one in which the underlying assumption is that missing data are conditionally MAR. When data are MNAR, it means that the missingness mechanism cannot be ignored, and must be modelled. Each of the MNAR models factorize the likelihood of the full data, f(Y,R|X), in different ways. For a recent review of MNAR models, see Ref. 69.

8.3 Pattern mixture models

The idea behind pattern mixture models is to allow parameters to vary according to missingness patterns. The estimates are then combined using weights based on the proportion of individuals in each pattern. Thus, the model is conditional upon membership in a particular dropout pattern. The likelihood factorization is $f(Y, R|X) = f(Y|R, X)f(R|X)$. Various pattern mixture models can be fit, both within the repeated measures and the growth curve frameworks.[70] These models are suited only to situations where the number of patterns is small or can be reasonably combined to give a small number. For data sets with n planned assessments, there are 2^n potential patterns where each assessment may or may not be observed. In most trials, there will be fewer patterns observed. The most frequently observed are patterns associated with dropout (e.g. once one assessment is missing, the following ones will also be missing). For the renal cancer data, there are 16 (2^4) potential patterns, however almost all ($> 95\%$) of the patients are not observed again after they miss their first assessment. Collapsing the few exceptions into alternative patterns based on the timing of the last observation yields 4 patterns, as shown in Figure 1.

8.3.1 Identifying restrictions

The primary difficulty with these models is that they are non-identifiable and all the parameters cannot be directly estimated from the observed data. For example, the means for the last time-points, as shown in Figure 1, cannot be estimated except for pattern 1, without making additional assumptions. These assumptions are known as identifying restrictions. One set of restrictions is (1) CCMV: Complete Case Missing Values, where information about the missing values is borrowed from the complete cases (pattern 1 from Figure 1); (2) ACMV: Available Case Missing Values, where all available data are used for imputing the means for the missing observations; and (3) NCMV: Neighbouring Case Missing Values, where available data from the nearest neighbour is used to impute means for the missing observations. CCMV (assumption 1) is the most restrictive assumption and roughly corresponds to there being no differences between patients who drop out (patterns 2–4 from Figure 1) and who complete (pattern 1) the trial. ACMV (assumption 2) will produce virtually the same results as the mixed models for monotone missing data. NCMV (assumption 3) is appropriate when conditional on the nearest neighbor, data are MAR. For example, pattern 3 is the neighboring pattern for missing assessments at the second time-point and pattern 1 is the neighboring pattern for missing assessments at the fourth time-point. It may not be too difficult to believe that estimates missing at the fourth assessment in pattern 2 are MAR when we only consider patients who complete 3 or 4 assessments (patterns 1 and 2), but it is more difficult to justify an assumption of MAR at the fourth assessment when we consider patients who complete only 1 assessment (pattern 4) pooled with those who complete all four assessments (pattern 1). Alternative restrictions include a variety of extrapolation techniques, including extrapolating the slope when trajectories are linear, the slope of the nearest neighbour, and using other polynomial

functions.[70] Multiple authors[71,72] have proposed sensitivity analyses using a penalty (delta) which is varied to examine the sensitivity of the results to such a penalty. Defining sensible extrapolations prior to observing the trajectories of each pattern can be difficult unless the investigators have had extensive prior experience with the population under study.

8.3.2 Standard errors

Because the estimates of the marginal means involve the product of two estimates (parameters and proportions), special methods need to be used to obtain the SEs including the delta method or bootstrapping.[19,32,73] If bootstrapping is used care should be taken that the longitudinal (non-independent) nature of the data is accounted for, by sampling patients rather than observations.

We illustrate using two approaches. We fit an RM pattern-mixture model and estimated SE with the bootstrap. Assuming an NCMV restriction, both the trajectories (Figure 2) and the estimate of the treatment differences (Table 4) are similar to the mixed model estimates which assumed data were MAR. Using an extrapolation of the last mean with an accruing penalty (delta) for each missed assessment, the estimates of treatment differences are also similar, but the within-group trajectories are sensitive to the assumed value of delta.

Table 4. Summary of estimates of the difference in quality of life (QoL) between experimental and control groups at the fourth time-point for the repeated measures models (RM, time is categorical) and at 17 weeks for the growth curve (GC, time is continuous) models.

Missingness assumption	Method	RM[a] or GC[b]	Estimate $QoL_{exp} - QoL_{ctl}$	SE	*p* Value
MCAR	*T*-test	-	−0.36	4.28	0.9
	ANCOVA	-	−2.22	3.50	0.5
	GEE	RM	−0.36	4.17	0.9
	GEE	GC	−1.91	3.07	0.5
MAR	Mixed model	RM	−8.08	3.67	0.03
	Mixed model	GC	−7.04	2.63	0.008
	Mixed model−MI[c]	RM	−11.87	5.20	0.02
	Unweighted GEE−MI[c]	RM	−11.87	5.20	0.02
	Weighted GEE	RM	−6.15	6.42	0.3
	Doubly robust GEE	RM	−9.67	14.01	0.5
	DR-GEE with truncated weights[d]	RM	−15.38	8.80	0.08
MNAR	Multivariate QoL & survival	RM	−8.20	3.59	0.02
	Multivariate QOL & log(time on treatment)	RM	−12.73	3.83	0.001
	Pattern mixture with NCMV	RM	−7.01	2.20	0.001
	Pattern mixture delta = 0		−6.30	2.43	0.01
	Pattern mixture delta = 5		−7.60	2.43	0.002
	Pattern mixture delta = 10		−8.90	2.43	0.0002

SE: standard error; GC: growth curve; MI: multiple imputation; GEE: generalized estimating equations; DR: doubly robust; MCAR: missing completely at random; MNAR: missing not at random; NCMV: neighbouring case missing values.

[a]GC models included terms for treatment, weeks (continuous), weeks^2, treatment × weeks, treatment × weeks.^{2}

[b]RM models included treatment, time (categorical), treatment, treatment × time.

[c]Imputation models for MI were performed by treatment, using log survival time, log time on treatment, and risk group.

[d]Replacing highest 1% of weights with 99th percentile.

8.4 Shared parameter models

In shared parameter models, time to dropout (or some other time to event, like survival time) or auxiliary variable and the longitudinal outcome are modelled "separately" but share common parameters (e.g. random effects[74–78]). Given the random effects, the two outcomes are assumed to be independent. The factorization is $f(Y, R|X) = \int [f(Y|b,X)f(R|b,X)\, dF(b|x)]$, where b is a subject's random effect. Wang and Hall[38] review these models when auxiliary data, rather than event time data, are used.

With our example, we attempted to fit several shared (random effects) parameter models but had difficulties. This illustrates a practical point in using these models: if there is little or no variation in one of the random effects then estimating the correlation is likely to fail and will not be an appropriate sensitivity analysis. In this study, there is minimal variation in the random slopes. Varying the starting parameters, increasing the number of quadrature points, and trying different distributions for the survival times did not result in convergence.

8.5 Joint multivariate models

Carpenter and Kenward[17] suggest an approach that can be used in the RM context where the outcome of interest and covariate(s) predictive of missingness (measured at or after baseline) are modelled in a joint multivariate model. We fit two models. The first used the log of survival time and simply included it as an outcome along with QoL. This treatment estimate was -8.20, $p = 0.02$, which is similar to several other model's estimates. Because survival time is a censored variable, and may not be appropriate, we also used the log of time on treatment and found an estimate of -12.73, $p = 0.001$.

8.6 Selection models

Selection models require a specification of the model for the full data (both observed and missing), the covariates, and the dropout probability. The factorization is $f(Y, R|X) = f(Y|X)f(R|Y, X)$. The model proposed by Diggle and Kenward[79] is the most well-know selection model for longitudinal data. It was a great step forward in the understanding of non-ignorable data, however, its usefulness has been minimal in practice. The model for $f(R|Y, X)$, as proposed, is a logistic model for dropout that depends on the covariates (X), the previously observed value of Y, and the current value of Y. Since the current value of Y is missing for all those who drop out, estimating this last piece of the dropout model assumes that one knows the distribution of the outcome; when this distribution is misspecified, the dropout model is incorrect and will not satisfy the assumptions of the selection model. This simple version of the model (depending only on the previous assessment) also requires at least two observations for each subject, which would exclude 10% of the patients in the renal cancer trial. These models require special software to run; Carpenter and Kenward give WinBUGS code.[17] Given these limitations (and others) we have not attempted to fit a selection model.

9 Conclusion

We have discussed various aspects of missing data in PROs, and demonstrated several analytical approaches with QoL data from an RCT comparing two therapies for renal-cell carcinoma, while emphasizing practical issues of implementation. The key issues when carrying out a longitudinal study with PROs are to minimize missing data through careful planning in the design and to perform

analyses which use all patient data while making reasonable (and examined) assumptions about missing data.

In the example, the wide range of estimates of treatment effect, from −0.35 to −15.38, underscores the sensitivity of the results in which different modelling approaches rely on different assumptions. It was also shown that within-group estimates of change over time are also sensitive to assumptions (see Figure 2).

MCAR methods, while simple, make strong assumptions about missing data that are rarely correct in studies of patients with morbidity and/or mortality. Perhaps researchers use simple imputation and MCAR approaches because they are wary of the assumptions that must be made for MAR methods. Their logic for using simple methods may be that all approaches involve assumptions, so why not use simple approaches? Qian et al.,[80] for example, state that they assumed the data were MCAR (despite dropout due to death of 13%) because "they wanted to keep the analyses manageable and it is not easy to identify the missing value processes in practice." However, the assumptions for an MAR analysis are not nearly as strong (or in some cases, such as LOCF, as unrealistic) as those required for an MCAR analysis. Furthermore, it has been shown that MAR analyses are fairly robust to data which are MNAR, unlike MCAR methods.[22] Simple imputation may be popular because of an attempt to follow the intention-to-treat principle.[81] However, MI and methods that assume data are MAR can be consistent with intention to treat.[22]

Our view, along with others in the missing data field,[17,21–23,57] is that a sensible approach is to base the primary analysis on MAR assumption and then to use MNAR models, if possible, for sensitivity analyses. Researchers and journal editors may be uncomfortable with more than one set of results, and results from the sensitivity analysis may have to go in the discussion and electronic appendices. However, one cannot be certain about results when there are substantial amounts of missing data without these analyses. Caution must be exercised in interpretation of results, and the inherent uncertainty in analysing patient-reported data where some are missing needs to be acknowledged.

Acknowledgement

Data presented are derived (and used with permission) from a trial conducted by Memorial Sloan-Kettering Cancer Center and Eastern Cooperative Oncology Group funded by the National Cancer Institute grant CA-05826.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Willke RJ, Burke LB and Erickson P. Measuring treatment impact: A review of patient-reported outcomes and other efficacy endpoints in approved product labels. *Contr Clin Trials* 2004; **25**: 535–552.
2. Au HJRJ, Brundage M, Palmer M, et al.; NCIC CTG Quality of Life Committee. Added value of health-related quality of life measurement in cancer clinical trials: the experience of the NCIC CTG. *Expert Rev Pharmacoecon Outcomes Res* 2010; **10**: 119–128.
3. Cella DF, Tulsky DS, Gray G, et al. The functional assessment of cancer therapy scale: Development and validation of the general measure. *J Clin Oncol* 1993; **11**: 570–579.
4. Hahn EA, Cella D, Chassany O, et al. Precision of health-related quality-of-life data compared with other clinical measures. *Mayo Clin Proc* 2007; **82**: 1244–1254.
5. Fielding S, Maclennan G, Cook JA, et al. 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; **9**: 51.
6. Wood AM, White IR and Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials* 2004; **1**: 368–376.
7. Jeličić H, Phelps E and Lerner RM. Use of missing data methods in longitudinal studies: the persistence of bad

practices in developmental psychology. *Developmental Psychol* 2009; **45**: 1195–1199.
8. Schlomer GL, Bauman S and Card NA. Best practices for missing data management in counseling psychology. *J Couns Psychol* 2010; **57**: 1–10.
9. Molnar FJ, Man-Son-Hing M, Hutton B, et al. Have last-observation-carried-forward analyses caused us to favour more toxic dementia therapies over less toxic alternatives? A systematic review. *Open Med* 2009; **3**: 1–20.
10. Gueorguieva R and Krystal JH. Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry. *Arch Gen Psychiatry* 2004; **61**: 310–317.
11. Tooth L, Ware R, Bain C, et al. Quality of reporting of observational longitudinal research. *Am J Epidemiol* 2005; **161**: 280–288.
12. Vach W and Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *Am J Epidemiol* 1991; **134**: 895–907.
13. Harris AHS, Reeder R and Hyun JK. Common statistical and research design problems in manuscripts submitted to high-impact psychiatry journals: What editors and reviewers want authors to know. *J Psychiatr Res* 2009; **43**: 1231–1234.
14. Harris AHS, Reeder R and Hyun JK. Common statistical and research design problems in manuscripts submitted to high-impact public health journals. *Open Public Health* 2009; **2**: 44–48.
15. Harris AHS, Reeder R and Hyun JK. Survey of editors and reviewers of high-impact psychology journals: statistical and research design problems in submitted manuscripts. *J Psychol* 2011; **143**: 195–209.
16. Chan AW and Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet* 2005; **365**: 1159–1162.
17. Carpenter J and Kenward M. *Missing data in randomised controlled trials - a practical guide*. Birmingham: National Institute for Health Research, 2008.
18. Group. IEEW. Statistical principles for clinical trials: ICH Harmonised Tripartite Guideline. *Stat Med* 1999; **18**: 1905–1942.
19. Fairclough DF. *Design and analysis of quality of life studies in clincial trials*, 2nd edn. Boca Raton, Fl: Chapman & Hall/CRC, 2010.
20. Mallinckrodt CH, Watkin JG, Molenberghs G, et al. Choice of the primary analysis in longitudinal clinical trials. *Pharmaceut Stat* 2004; **3**: 161–169.
21. Molenberghs G and Kenward MG. *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons, 2007.
22. Molenberghs G, Thijs H, Jansen I, et al. Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 2004; **5**: 445–464.
23. Fitzmaurice GM, Laird NM and Ware JH. *Applied longitudinal analysis*, 2nd edn. Hoboken NJ: Wiley, 2011.
24. Molenberghs G and Verbeke G. *Linear Mixed Models for Longitudinal Data*. 2nd ed. New York: Springer, 2009.
25. Motzer RJ, Murphy BA, Bacik J, et al. Phase III trial of interferon alfa-2a with or without 13-cis-retinoic acid for patients with advanced renal cell carcinoma. *J Clin Oncol* 2000; **18**: 2972–2980.
26. Webster K, Cella D and Yost K. The Functional Assessment of Chronic Illness Therapy (FACIT) measurement system: Properties, applications, and interpretation. *Health Qual Life Outcomes* 2003; **1**: 79.
27. Kenward MG and Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**: 983–997.
28. Dempster A, Laird N and Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc Series B* 1977; **39**: 1–38.
29. Fairclough DL and Cella DF. Functional assessment of cancer therapy (FACT-G): Non-response to individual questions. *Qual Life Res* 1996; **5**: 321–329.
30. Fayers PM, Curran D and Machin D. Incomplete quality of life data in randomized trials: Missing items. *Stat Med* 1998; **17**: 679–696.
31. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–592.
32. Hogan JW, Roy J and Korkontzelou C. Tutorial in biostatistics. Handling drop-out in longitudinal studies. *Stat Med* 2004; **23**: 1455–1497.
33. Frangakis CE and Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; **58**: 21–29.
34. Kurland BF, Johnson LL, Egleston BL, et al. Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. *Stat Sci* 2009; **24**: 211–222.
35. Osoba D, Bezjak A, Brundage M, et al. Evaluating health-related quality of life in cancer clinical trials: The National Cancer Institute of Canada Clinical Trials Group experience. *Value Health* 2007; **10**: S138–S145.
36. Sloan JA, Dueck AC, Erickson PA, et al. Analysis and interpretation of results based on patient-reported outcomes. *Value Health* 2007; **10**: S106–S115.
37. Collins LM, Schafer JL and Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Meth* 2001; **6**: 330–351.
38. Wang C and Hall CB. Correction of bias from non-random missing longitudinal data using auxiliary information. *Stat Med* 2010; **29**: 671–679.
39. Ibrahim JG, Lipsitz SR and Norton N. Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *J R Stat Soc Ser C Appl Stat* 2001; **50**: 361–373.
40. Baker SG, Fitzmaurice GM, Freedman LS, et al. Simple adjustments for randomized trials with nonrandomly missing or censored outcomes arising from informative covariates. *Biostatistics* 2006; **7**: 29–40.
41. Pocock SJ, Assmann SE, Enos LE, et al. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Stat Med* 2002; **21**: 2917–2930.
42. Committee for proprietary medicinal products (CPMP). points to consider on adjustment for baseline covariates. *Stat Med* 2004; **23**: 701–709.
43. Friedman LM, Furberg CD and Demets DL. *Fundamentals of clinical trials*, 4th edn. New York: Springer, 2010.
44. Piantadosi S. *Clinical trials:A methodologic perspective*, 2nd edn. Hoboken: Wiley, 2005.
45. Donaldson GW and Moinpour CM. Learning to live with missing quality-of-life data in advanced-stage disease trials. *J Clin Oncol* 2005; **23**: 7380–7384.
46. Robins JM, Rotnitzky A and Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Stat Assoc* 1995; **90**: 106–121.
47. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; **340**: c869.
48. Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.
49. Burton P, Gurrin L and Sly P. Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med* 1998; **17**: 1261–1291.

50. Hanley JA, Negassa A, Edwardes MDDB, et al. Statistical analysis of correlated data using generalized estimating equations: An orientation. *Am J Epidemiol* 2003; **157**: 364–375.
51. Lipsitz SR, Molenberghs G, Fitzmaurice GM, et al. GEE with Gaussian estimation of the correlations when data are incomplete. *Biometrics* 2000; **56**: 528–536.
52. Birhanu T, Molenberghs G, Sotto C, et al. Doubly robust and multiple-imputation-based generalized estimating equations. *J Biopharm Stat* 2011; **21**: 202–225.
53. Fairclough DL. Summary measures and statistics for comparison of quality of life in a clinical trial of cancer therapy. *Stat Med* 1997; **16**: 1197–1209.
54. Matthews JNS, Altman DG, Campbell MJ, et al. Analysis of serial measurements in medical research. *Br Med J* 1990; **300**: 230–235.
55. Bell ML, Fairclough DL and King MT. Bias in area under the curve for quality of life longitudinal clinical trials with missing data: comparison of summary measures versus summary statistics. In submission.
56. Molenberghs G and Verbeke G. *Linear mixed models for longitudinal data*, 2nd edn. New York: Springer, 2009.
57. Mallinckrodt CH, Clark WS, Carroll RJ, et al. Assessing response profiles from incomplete longitudinal clinical trial data under regulatory considerations. *J Biopharm Stat* 2003; **13**: 179–190.
58. Schafer JL. Multiple imputation: A primer. *Stat Methods Med Res* 1999; **8**: 3–15.
59. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons, 1987.
60. Horton N and Kleinman K. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat* 2007; **61**: 79–90.
61. Nevalainen J, Kenward MG and Virtanen SM. Missing values in longitudinal dietary data: A multiple imputation approach based on a fully conditional specification. *Stat Med* 2009; **28**: 3657–3669.
62. Schafer JL and Graham JW. Missing data: Our view of the state of the art. *Psychol Meth* 2002; **7**: 147–177.
63. Bang H and Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**: 962–972.
64. Carpenter JR, Kenward MG and Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J Royal Stat Soc Ser A: Stat Soc* 2006; **169**: 571–584.
65. Seaman S and Copas A. Doubly robust generalized estimating equations for longitudinal data. *Stat Med* 2009; **28**: 937–955.
66. Hosmer D and Lemeshow S. Goodness of fit tests for the multiple logistic regression model. *Commun Stat, Part A-Theory Meth* 1980; **9**: 1043–1069.
67. Kang JDY and Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci* 2007; **22**: 523–539.
68. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education*. Washington DC: National Academies Press, 2010.
69. Ibrahim JG and Molenberghs G. Missing data methods in longitudinal studies: A review. *Test* 2009; **18**: 1–43.
70. Demirtas H and Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Stat Med* 2003; **22**: 2553–2575.
71. Diehr P, Patrick D, Hedrick S, et al. Including deaths when measuring health status over time. *Med Care* 1995; **33**(supp): AS164–AS172.
72. van Buuren S, Boshuizen HC and Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999; **18**: 681–694.
73. Curran D, Molenberghs G, Aaronson NK, et al. Analysing longitudinal continuous quality of life data with dropout. *Stat Methods Med Res* 2002; **11**: 5–23.
74. Wu MC and Bailey KR. Analyzing changes in the presence of informative right censoring caused by death and withrawal. *Stat Med* 1988; **7**: 337–346.
75. DeGruttola V and Tu XM. Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics* 1994; **50**: 1003–1014.
76. Elashoff RM, Li G and Li N. An approach to joint analysis of longitudinal measurements and competing risks failure time data. *Stat Med* 2007; **26**: 2813–2835.
77. Thiébaut R, Jacqmin-Gadda H, Babiker A, et al. Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+cell count and HIV RNA viral load in response to treatment of HIV infection. *Stat Med* 2005; **24**: 65–82.
78. Vonesh EF, Greene T and Schluchter MD. Shared parameter models for the joint analysis of longitudinal data and event times. *Stat Med* 2006; **25**: 143–163.
79. Diggle PJ and Kenward MG. Informative drop-out in longitudinal data analysis. *J Royal Stat Soc Ser C (Appl Stat)* 1994; **43**: 49–93.
80. Qian W, Parmar MKB, Sambrook RJ, et al. Analysis of messy longitudinal data from a randomized clinical trial. *Stat Med* 2000; **19**: 2657–2674.
81. Altman DG. Missing outcomes in randomized trials: Addressing the dilemma. *Open Med* 2009; **3**: e51–e53.