JOURNAL OF BIOPHARMACEUTICAL STATISTICS
Vol. 14, No. 1, pp. 53–71, 2004

The Analysis, Interpretation, and Presentation of Quality of Life Data

Richard Stephens*

Cancer Division, Medical Research Council Clinical Trials Unit, London, UK

ABSTRACT

All too often in clinical trials the assessment of quality of life is seen as a bolt-on study. Consequently insufficient consideration is often given to its design, collection, analysis and presentation, and its impact on the trial results and on clinical practice is minimal. In many trials quality of life is a key endpoint, and it is vital that quality of life expertise is involved as soon as possible in the design. Setting a priori quality of life hypotheses will focus the decisions regarding which questionnaire to use, when to administer it, the sample size required, and the primary analyses. Nevertheless quality of life data are complex, and require much skill in determining how to deal with multi-dimensional and longitudinal data, much of which is often missing. There are no agreed standard ways of analysing and presenting quality of life data, but there are guidelines, which if followed, will add transparency to the way results have been calculated. Understanding the impact of treatments on their quality of life is vital to patients, and it is up to us, as statisticians and trialists, to present the data as clearly as we can.

Key Words: Quality of life; Hypotheses; Compliance; Palliation; Missing data; Longitudinal data; Multidimensional data.

*Correspondence: Richard Stephens, Cancer Division, Medical Research Council Clinical Trials Unit, 222 Euston Road, London NW1 2DA, UK; Fax: +44 20 7670 4818; E-mail: rs@ctu.mrc.ac.uk.

DOI: 10.1081/BIP-120028506 1054-3406 (Print); 1520-5711 (Online)

INTRODUCTION

Historically, the assessment of quality of life (QOL) has often been seen as an additional substudy to a clinical trial where the primary outcome is usually survival. This has often led to statisticians being presented with a large quantity of QOL data at the end of a trial and being asked to analyze and interpret the results. This can present numerous problems as there are no universally recognized methods of analysis for biomedical data, and with the multitude of potential comparisons, can lead to trawling through the data to find an interesting result, which of course may be completely spurious.

To avoid this dangerous scenario, QOL experts and statisticians need to be involved from the very conception of the trial to ensure that QOL assessment is appropriately included and that a priori hypotheses are formulated. Of course, this might result in the assessment of QOL not being included in some trials, which might save enormous amounts of time and effort. However, similarly, making a decision that the assessment of QOL is appropriate and setting hypotheses is also likely to make this a cost-effective exercise as it should ensure that a suitable QOL sample size is defined, that the correct questionnaires are used at the correct timepoints, and that the appropriate analyses are stated to answer the QOL question set.

QUALITY OF LIFE HYPOTHESES

When a priori hypotheses have been generated, analysis and interpretation can potentially become straightforward as everything is set down in the protocol, compliance can be monitored, and interim analyses (for the purposes of independent data monitoring committees) tested throughout the trial.

Problems occur of course when a priori hypotheses have not been set and when compliance has not been monitored and chased, resulting in a dataset with much missing data. Even in this situation there is an option to generate some retrospective hypotheses, by surveying clinicians, nurses (Groenvold and Fayers, 1998), and/or patients to clarify the important QOL domains, the key timepoints, and the degree of difference that would need to be seen to make a significant change in clinical practice.

For those not familiar with QL hypotheses, it is worth presenting a couple of examples.

In a United Kingdom Medical Research Council (MRC) trial comparing oral chemotherapy with standard intravenous chemotherapy for patients with small cell lung cancer (Medical Research Council Lung Cancer Working Party, 1996a), no difference was expected in overall survival, but it was hoped that oral chemotherapy would achieve equivalent palliation of key symptoms. The primary end point of the trial was therefore formulated as "oral chemotherapy should achieve at least equivalent palliation of major symptoms (a reduction in the sum of severity scores for cough, pain, anorexia, and shortness of breath from baseline to three months)."

A more complex algorithm was designed for a trial of two chemotherapy drugs for patients with pancreatic cancer (Burris and Storniolo, 1997). In order to divide

patients into those who "responded" to the therapies and those who did not, a number of aspects were assessed and combined, as shown in Fig. 1. This is a very novel approach and it is perhaps surprising that it has not been taken up in other trials. Perhaps it is considered to be too complex, but putting effort into designing such a hypothesis at the start of a trial will probably save much more time and effort at the end.

Finally, in the Big Lung Trial (Stephens et al., 2002), which compared chemotherapy and no chemotherapy for all patients with non-small cell lung cancer, no a priori hypotheses were set up. However, before the end of the trial, clinicians on the Trial Management Group were asked to consider what changes in which QOL domains and items would be sufficient to make them consider changing their practice. Although not an ideal solution, it was felt that this would focus the analysis. The result of this survey was that the comparison of global QOL at 12 weeks from randomization was considered the primary end point, and secondary end points were identified as the levels of pain, dyspnoea, fatigue, and emotional and physical functioning at 12 weeks.

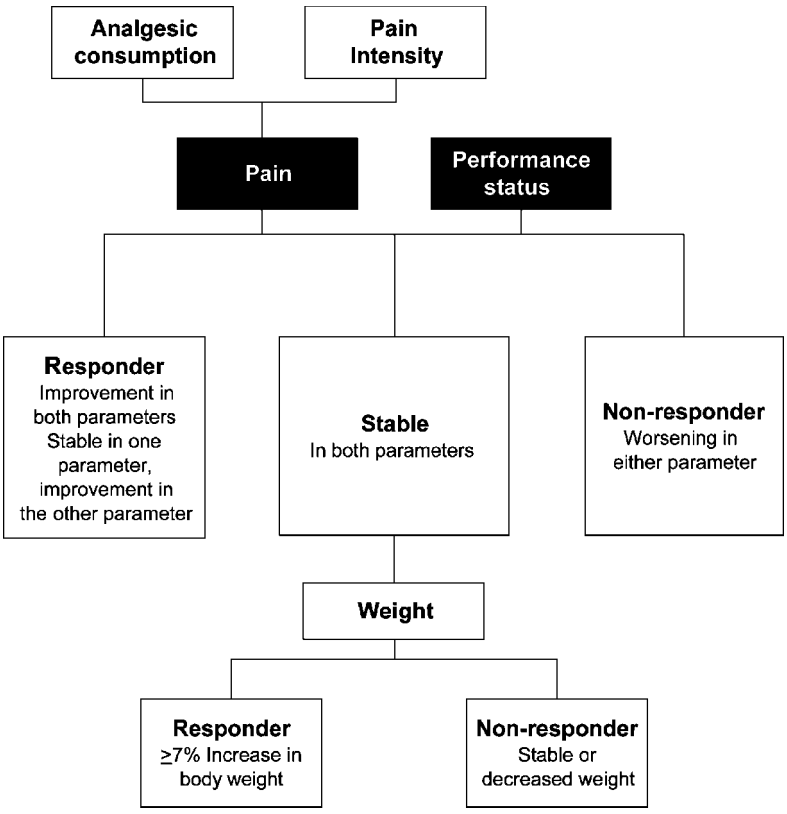

Figure 1. Definition of a QOL responder in a trial of two chemotherapy drugs for patients with pancreatic cancer (adapted from Burris and Storniolo, Eur. J. Cancer 1997, 33, suppl 1, S18–S22).

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Sample Size

Setting clear QOL hypotheses will allow a sample size to be calculated. However, one cannot assume of course that the sample size set for, say, survival will be sufficient for QOL comparison. This would be especially true if a new treatment was expected to result in improved survival but equivalent QOL. In addition, one needs to take into account the likely level of missing data (forms not completed by patients at the right time) and attrition (patient dropout, due usually in cancer trials to death).

GUIDELINES FOR ANALYSIS

Before we consider the analysis options for QOL data, it is worth reiterating two key messages. First, that the primary analysis should always be an intention-to-treat analysis and, second, that subgroup analysis should always be treated with scepticism.

Intention-to-Treat

The aim of most randomized trials is to compare the treatment of patients with policy A vs. policy B. If, for instance, policy A is a complex treatment and only half the patients actually receive it, that is a key message from the trial, and removing these patients from any analysis will give a false impression of the efficacy of that treatment. Future groups of patients will all be subject to the same rates of success and failure.

Patients are often removed from analyses for a variety of reasons, but all of these seemingly valid reasons can induce bias:

- Patients may not receive any of their allocated treatment, perhaps because they change their mind or their condition deteriorates. This is likely to be highly related to the treatment given, for example, if one of the arms is no treatment, then no patients could be removed from this arm.
- Patients may receive some, but not all, of their treatment. Again this may relate to the treatment, for example, many patients may not complete long courses of chemotherapy.
- Patients may not be assessed. Before an assessment point for, say response, is reached, patients may die or be lost to follow-up. Removing them from the analysis will bias the result, as it will only be based on the compliant surviving patients. To get round this, ''deaths'' can be treated as ''failures'' or alternative analyses found that take into account censored data.
- Further tests after randomization may indicate patients are ineligible. This may only occur in one arm of the trial, as patients, for example, may need extra investigations prior to surgery. Removing such patients would clearly bias the analysis.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

- Independent review may decide patients are ineligible. Although it seems extremely logical to exclude such patients, trials usually are trying to assess how effective a treatment will be in routine practice. It is unlikely then that expert independent review will be undertaken routinely.

The one exception to the rule of intention-to-treat is equivalence or noninferiority trials, in which using intention-to-treat may tend to dilute any effect between the treatments, making them appear more similar than they actually are.

Baseline

The baseline for all patients should always be the date of randomization. This is the one common timepoint from which everything can be measured and compared. Using start (or end) of therapy as baseline can potentially introduce many of the problems listed above.

Subgroup Analysis

Unless prespecified, all subgroup analyses should be considered as exploratory. As there will be an almost indefinite number of possible subgroups, the chance of finding a significant result increases as the number of comparisons increases. Trials are rarely powered for detecting such differences. Differences in different subgroups are more likely if there is an overall difference, as we are looking for quantitative differences between subgroups. If there is a difference in subgroups, it is important to consider whether this is plausible (e.g., is there a trend with increasing age or is there external evidence showing the same thing).

ANALYSIS OF QOL DATA

When there are no a priori hypotheses, or even when there are and additional exploratory analyses need to be conducted, the very nature of QOL data means that there are a number of aspects that have to be considered and decisions taken:

- There will inevitably be missing data—can it be imputed or analyses used which account for it?
- The data is multidimensional—can individual items be analyzed and presented or, to simplify the data, can items be combined into domains and subscales?
- The data is longitudinal, i.e., collected over time—can this be presented as such or are methods of summarizing the data required?

Any of the above three aspects is difficult to deal with, combined they make the analysis of QOL data formidable.

In this paper we will consider each of these three concerns individually and then look at possible options for combining them in an attempt to assess palliation.

Missing Data

It is extremely difficult to attain good QOL compliance, especially in multicenter trials with patients who are often rapidly deteriorating. Compliance, the number of forms completed at the correct timepoint as a proportion of the number expected, has been reported to range from 23% to 96% in various trials (Hurny et al., 1992). Sadura et al. (1992) showed that compliance of over 95% was possible given adequate resources, running pretrial workshops for local QOL coordinators to emphasize the need for QOL in the trial, and constantly monitoring and feeding back information on compliance to individual centers. It is clear in most trials that the problem with compliance is not the patient but the lack of commitment in the local center.

The best answer to the missing data question is not to have any! Missing data may severely restrict the analyses that can be done, and any conclusions based on subgroups of patients with full data may not be generalizable even to the rest of the patients in the trial, let alone future patients.

Compliance

It is important to present information on patients' levels of compliance in completing the questionnaires. This should take into account that questionnaires have been completed at the specified timepoints and of course to check that any findings are consistent across all arms of the trial.

It is unlikely that patients who complete all QOL forms will be truly representative of the whole group of patients, for example, they have to be survivors. This need not affect the between-treatment comparison, as it is likely that the reasons for noncompliance will be the same across all arms.

Hopwood et al. (1994) defined compliance as the number of forms completed as a proportion of those expected, expected being the number of patients alive at the specified timepoint. They also suggested the use of acceptable time windows around the specified timepoints to allow for some variation in patient treatment and follow-up. These might be narrow around key timepoints, wider for later timepoints, symmetrical, or asymmetrical (to allow for drift in follow-up assessment). Table 1 shows compliance in the completion of the hospital anxiety and depression scale (HADS) (Zigmond and Snaith, 1983) from an MRC trial of four drugs vs. two drugs for small cell lung cancer (Medical Research Council Lung Cancer Working Party, 1996b).

It is inevitable that some data will be missing, and most often it will be missing from the most ill patients, the very patients on whom it is most important to collect QOL data. Data can be classed as missing completely at random (MCAR) when it is not connected to any factor, missing at random (MAR) when it is unconnected with patient characteristics but related perhaps to, say, a particular center, or not missing at random (NMAR) when it is missing for a reason connected to a patient's condition, usually poor performance status. The fact that QOL data is nearly always NMAR makes it extremely difficult to impute, as the standard method for

Table 1. Compliance in the completion of HADS forms in an MRC trial of chemotherapy for small cell lung cancer.

Assessment	Time due (day)	Time window (days)	Patients		Forms received
			Dead	Alive (forms expected)	
1	Baseline (day 0)	−7 to +1	0	310	237 (76%)
2	Second cycle (day 21)	±14	58	252	166 (66%)
3	Third cycle (day 42)	±14	73	237	138 (58%)
4	Month 3 (day 91)	±28	116	194	82 (42%)
5	Month 4 (day 122)	±28	140	170	64 (38%)

imputing missing data, extrapolating in some way from within (or outwith) the data set, cannot be used. It would be illogical to impute data for patients who are unwell and unable to complete forms from those who are relatively fit and well and have completed forms.

As opposed to missing forms, very few papers report the extent of missing items. Although on average patients fail to complete, or choose not to complete, only a tiny proportion of questions (Fairclough and Cella, 1996), this can still add up when patients are expected to complete long forms at numerous timepoints.

Most standard questionnaire manuals suggest methods of dealing with missing items (Cella, 1997; De Haes et al., 1996; Fayers et al., 1995). For example, if a patient has answered three of four questions relating to anxiety in a positive way, it is reasonable to assume that they would have answered the fourth in the same manner. Thus when calculating a score for the anxiety domain, the missing answer can be extrapolated from the other three. Most questionnaire manuals suggest that if 50% of the items in a scale are available, the subscale score can be calculated. However, this can still be a risky thing to do, and Fayers et al. (1998) list a number of checks that should ideally be performed before any imputation of items.

Of course, much "missing" data may be from patients who have died, and whereas in some circumstances this may be easily solved by allotting them the worst score (for example, the Karnovsky scale includes "dead" as a category), it is completely illogical to state that all dead patients have "severe cough" or are considered to be a "clinical case of depression".

If properly carried out, imputation can restore balance to the data and permit simpler analyses. Nevertheless, it is important to remember that any interpretation of the results in the presence of an incomplete dataset are not as convincing as the interpretation based on a complete dataset.

Graphical Summaries

A good starting point in getting a feel for the data may be, as suggested by Machin and Weeden (1998), to simply plot the scores from a QOL questionnaire against the time from randomization. Figures 2a and 2b shows the anxiety subscale score from the HADS questionnaire (Zigmond and Snaith, 1983) for the two regimens in the MRC small cell lung cancer trial (Medical Research Council Lung

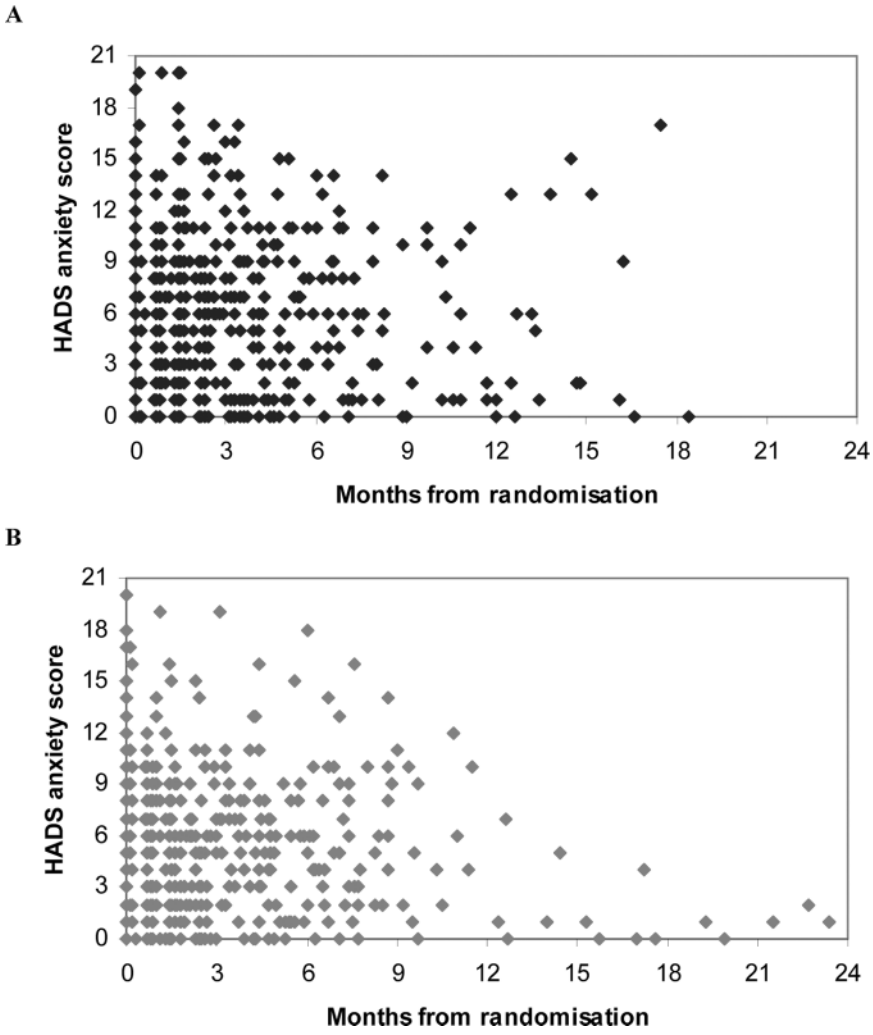

Figure 2. Anxiety subscale scores from the HADS questionnaire plotted against time from randomization in patients receiving (A) four drugs and (B) two drugs in an MRC SCLC trial. (*View this art in color in www.dekker.com.*)

Cancer Working Party, 1996b). This effectively shows the increasing amount of missing data over time, the drift from scheduled assessment points, and also the increasing proportion of patients reporting lower scores (i.e., "normal" levels of anxiety). However, whether the latter is a genuine improvement or merely an artefact of the fact that all the patients with baseline anxiety have not completed questionnaires at later timepoints cannot be deduced from this plot.

It can also be useful to plot individual patient profiles, and Fig. 3 shows the patient reported scores from the Rotterdam symptom checklist (De Haes et al., 1996) for "lack of appetite" for 12 patients (6 receiving continuous hyperfractionated

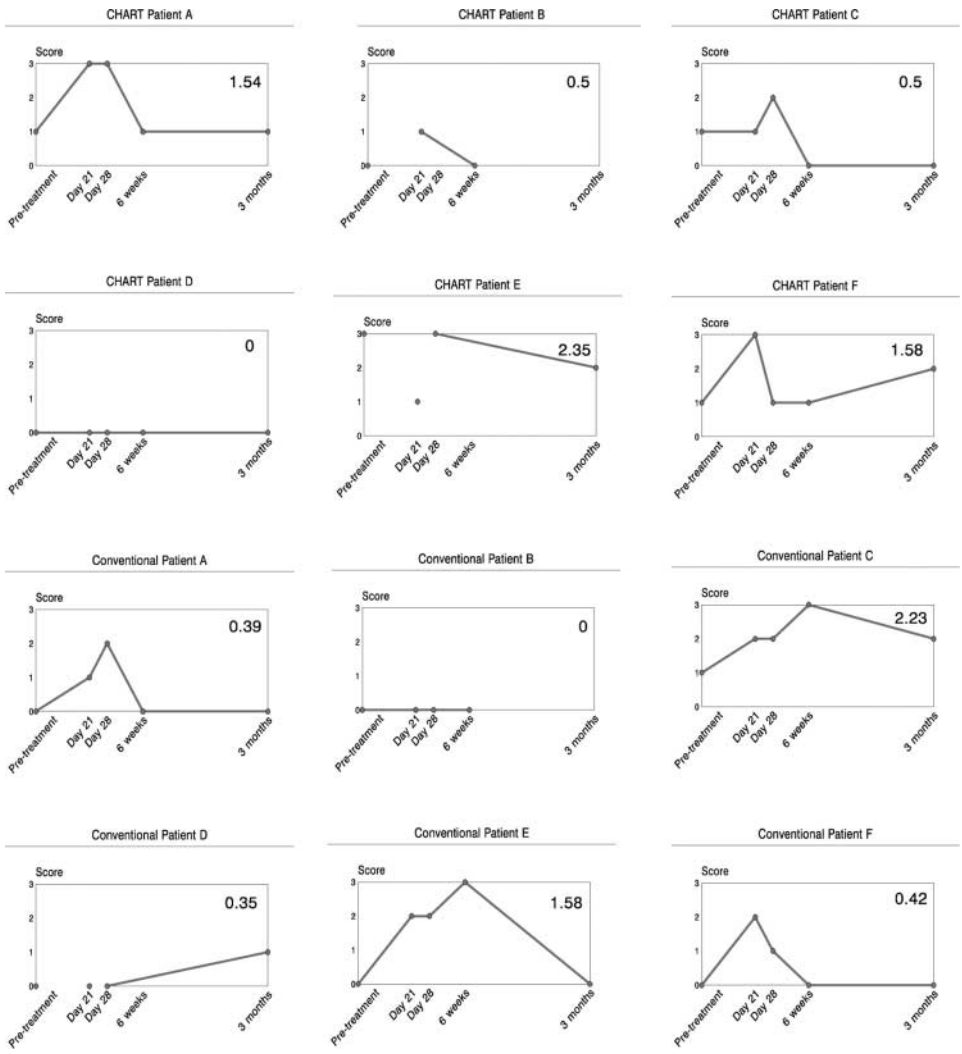

Figure 3. Individual patients "lack of appetite" scores over time from the CHART trial. (*View this art in color in www.dekker.com.*)

accelerated radiotherapy (CHART) and 6 conventional radiotherapy) in a lung cancer trial (Saunders et al., 1997). This shows the variability both within and between patients and also shows that missing data are common.

Multidimensional Data

Most standard QOL questionnaires include 30 to 40 individual items. While this number may be necessary to cover the range of symptoms, side effects and domains

of QOL, it is almost always too many to analyze and present. The common solution is to combine items into validated groups, domains, or a total score. In some scenarios, this is clearly the best option. For example, the HADS questionnaire (Zigmond and Snaith, 1983) includes seven separate questions, that relate to anxiety, each with four possible responses (scored 0–3). These seven questions are designed to be combined to produce a range of scores from 0 to 21, which, in turn, can be subdivided into three states (normal [scores 0–7], borderline [scores 8–10], and probably clinical case [scores 11–21]). This would be almost impossible to do with one four-response question, because of the complexity of QOL domains.

In other situations, it is suggested that all response scores relating to, say, physical symptoms are combined (De Haes et al., 1996). This is much more contentious, as, for example, chemotherapy for lung cancer may palliate presenting symptoms (cough, shortness of breath, haemoptysis) but cause side effects (nausea, vomiting, alopecia). Simply adding all these scores together may suggest no change in overall physical symptoms, which might in one sense, be accurate but equally totally uninformative about the real effect of treatment.

One suggested solution is to analyze by the domains first and, where differences are seen, to then analyze those domains by the individual items. However, this would not work in the example given above.

Longitudinal Data

QOL data are often collected for extended time periods, which in advanced cancer can often mean the remainder of a patient's life, as it is important to investigate not only the severity and duration of short-term transient effects but also any long-term and/or chronic problems with treatment. It is also important to know, in certain situations, whether patients return to "normal" life following cancer and its treatment. However, analyzing and comparing groups of patients with varying amounts of data, due to attrition, are complex. Nevertheless, having collected detailed longitudinal data, it would seem illogical to then reduce it to a summary score, although, as we shall see later, this may be the only way to obtain a statistical comparison.

Longitudinal data are often simply presented in a graphical format, which on occasions can be extremely informative. In an MRC trial of two radiotherapy schedules (39Gy/13f and 17Gy/2f) for the treatment of patients with inoperable non-small cell lung cancer (Medical Research Council Lung Cancer Working Party, 1996c), data were collected using daily diary cards (Fayers et al., 1991) on which patients were asked to report the severity of a number of symptoms and QOL aspects each evening. The resulting plot (Fig. 4) of the proportion of patients reporting dysphagia each day clearly showed that the longer radiotherapy schedule caused this side effect to be more severe, and of longer duration.

However, such descriptive plots can potentially be very misleading. An example of this is Fig. 5a where the mean "activity of daily living" score for a group of patients has been plotted over time (a low score indicates more difficulties with walking, shopping, etc., and a high score indicates less problems). The plot appears to show that the activity level of this group of patients is improving.

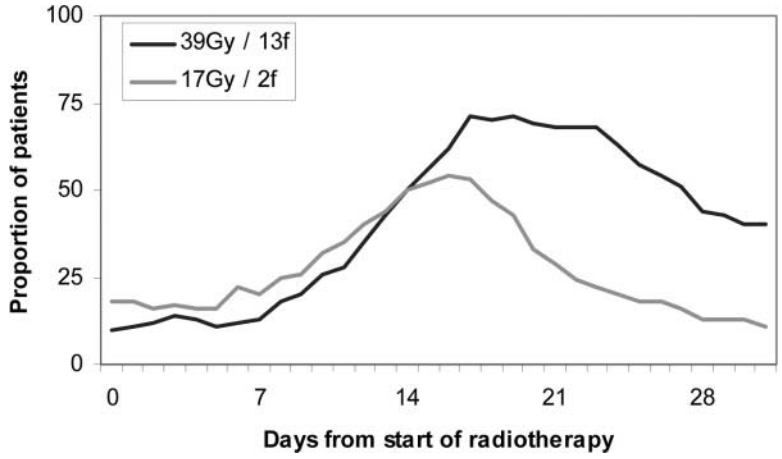

Figure 4. Proportion of patients reporting dysphagia on daily diary cards in an MRC radiotherapy trial. (*View this art in color in www.dekker.com.*)

However, if the group of patients is subdivided by the number of questionnaires returned and then plotted for each subgroup (Fig. 5b), it is clear that the presumed overall improvement is simply due to the progressive dropout of patients with lower scores (Hopwood et al., 1994). Plots such as Fig. 5a and their misinterpretation are common in the literature but can be clarified somewhat by stating the number of patients contributing to the plot at each timepoint along the x axis.

An alternative is to only plot the scores from patients with complete datasets, but this of course will undoubtedly greatly reduce the sample size. Also, as the data will have come from patients who are survivors and compliers, the result may not be generalizable to the whole trial population.

Summary Scores

An effective way of dealing with longitudinal data is to derive a summary score for each patient. Possible options could be the worst score reported, the best score, the mean, the median, the area under the curve (AUC), or the last available score. The advantage of using a summary score is that analyses are focused, statistics are valid, missing data can be accommodated, and the calculation is relatively straightforward and understandable to readers.

However, as there are numerous ways in which data can be summarized, it is essential that consideration is given to the summary measure chosen. One might chose the worst score in a trial evaluating a new less toxic treatment (although this would take no account of duration) or the best score where the aim is to palliate baseline symptoms. The pattern of change over time may also be a useful guide as to the choice of the best summary score. Thus, for an increasing or decreasing line, the regression coefficient, final value, or time to a certain value might be appropriate. Whereas for a line that peaks, the maximum or time to maximum might be chosen.

Note that the use of the mean or median as a summary score can often mask a proportion of patients with severe problems. An example of an inappropriate

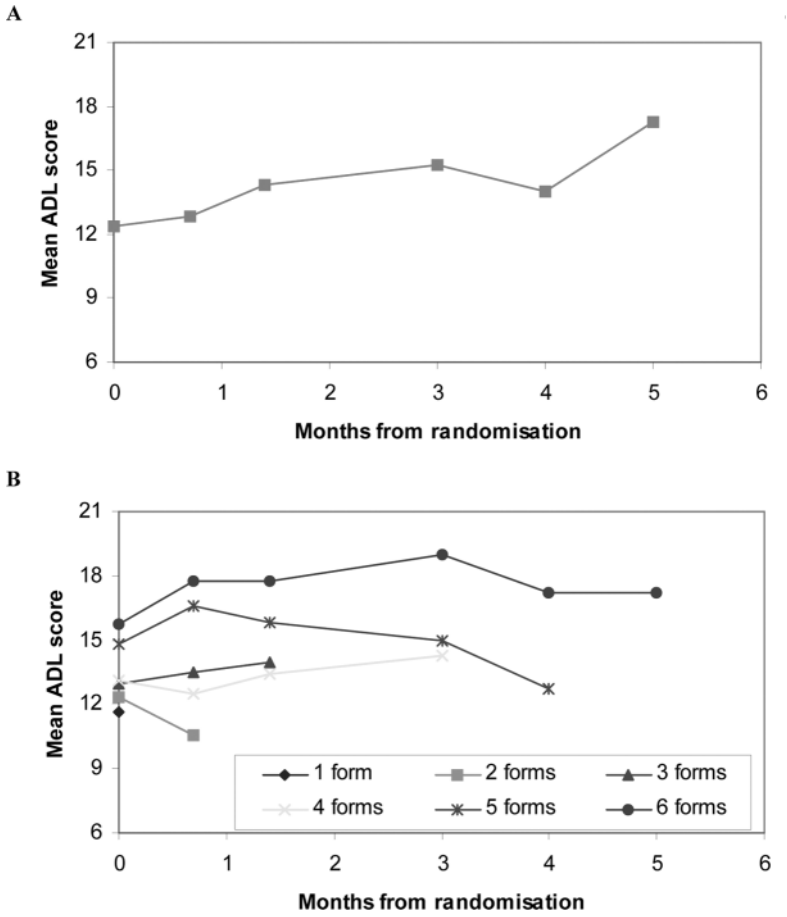

Figure 5. Mean "activity of daily living" scores. (*View this art in color in www.dekker.com.*)

summary might be the use of the mean anxiety score. Using the data from Fig. 2, the mean score at all timepoints for the total group of patients in each arm would be <8, but quoting such a mean score might easily infer that all patients report normal levels of anxiety (defined as a score of 0–7 for an individual patient) when this clearly is not the case.

Rather than plotting the mean or median score over time, plotting the proportion of patients who fall into a particular category (or categories) may be a more logical approach. For instance, in the example above, plotting the proportion of patients with baseline or case anxiety at each timepoint.

The choice of summary requires a good understanding of the patterns of change likely to be seen, and thus a pilot or feasibility study may be an excellent opportunity to collect this sort of information. For instance, two treatments may both cause nausea and vomiting, but the duration may be much longer in one treatment than in the other. This might then indicate the need to collect and compare data on duration as well as severity.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Area under the Curve

The calculation of the AUC seems like the most logical way to summarize each individual patient's total experience of a symptom, side effect, or QOL domain. In Fig. 3, the AUC summary score for each patient's lack of appetite is shown in the corner of each plot. In this example, the scores will range from 0 for a patient who reports no lack of appetite for the whole time period to 3 where the patient reports severe lack of appetite at every assessment. As with all summaries, there are disadvantages with the AUC, for example, assumptions have to made that there is a linear change between assessments: a patient experiencing a transient severe episode may have the same summary score as one with a chronic mild problem and, as with all summary scores, the ability to look at the pattern of change over time is lost.

Nevertheless, the AUC probably represents the best summary score, and once decisions have been made about how to deal with missing data, the treatment groups can be compared using the Wilcoxon or Mann-Whitney nonparametric tests. The paper by Bailey et al. (1998) gives a good practical example of the use of the AUC in a randomized trial.

As there is no agreed standard way of analyzing QOL data, it is often useful to conduct a number of different types of analysis to ensure that any conclusions are not just a function of the approach chosen. In the analysis of longitudinal QOL data, Qian et al. (2000) compared a variety of summary scores as well as a complex model-based approach to assess the sensitivity of any results. They showed that for this dataset at least, the analysis of simple summary scores produced similar results to that of more complex methods. However, by analyzing the data in a variety of ways they were able to reject inconsistent results that might otherwise have been wrongly emphasized.

Time to, or Duration of, an Event

Time-to-event analyses are frequently used in trials, usually of course to compare survival, but they can equally be used in the assessment of QOL to assess the time to, say, the worst, or best, score. However, censoring may be more of an issue in this situation. In survival analyses, surviving patients are censored at the last time seen, and it is assumed that they will have the same survival as other patients who have survived to this point in time. If we try to plot the time to improvement of a symptom, is it reasonable to assume that patients who die without improvement can be censored at the time of death and the assumption made that had they lived they too would have been palliated? Such an assumption could lead to a treatment that, say, causes 50% early deaths but palliates 25% of the survivors, appearing to be more palliative than one that causes no early deaths and palliates, say, 45% of all patients.

COMPLEX MODELS

Most of the more complex analyses require fitting a mathematical model to the data. The two most common approaches are multivariate analysis of variance (MANOVA) or mixed model (multilevel) models. Model-based analyses are theoretically more efficient than all the analyses presented above because they explicitly allow for correlation of repeated measures by including a covariance structure. Adjustments

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

for other possible variables can be made and time or time-dependent variables can also be included so that time-changing patterns can be considered. However, model-based analyses make a number of assumptions and pose a large number of restrictions. For example, MANOVA models only use complete cases, thus, some form of imputation is required; mixed models assume data are missing at random.

INTEGRATION OF SURVIVAL AND QOL

The concept of combining QOL and survival into a single statistic is very appealing, as it appears to overcome the subjective balancing of quality and quantity of life. An argument against simple survival analysis is that patients can only be classified in one of two states, alive and dead. In this situation, a patient who is bed-ridden with severe symptoms and a patient who is active and asymptomatic are both categorized as simply "alive." Quality adjusted life years (QALYs) are calculated on the basis that a value (called a utility) between 0 (dead) and 1 (fully fit) can be assigned to various intermediate health states (Kaplan, 1993). The time spent in each state can then be multiplied by this value to express survival in terms of QALYs. It is important to consider who provides the utility rating (patient, doctor, society, healthy individuals, etc.) and whether utilities change over time. Sensitivity analyses, using different utilities, may be necessary to confirm results.

Clinicians and trialists have not in the main embraced QALYs, perhaps because there is a feeling that this is an oversimplification and there is still a desire to consider all the various domains and items of QOL separately in order to make informed decisions about treatments.

An alternative way of combining QOL and survival is TWIST (time without symptoms and toxicity), which is a summation of the survival time during which patients report no symptoms or toxicity (Gelber and Goldhirsch, 1986; Goldhirsch et al., 1989). It is calculated by subtracting from the overall survival time periods when symptoms or other clinical events were present. This requires defining what events, or what severity of events, are relevant and what the time penalty for each should be.

PALLIATION

Commonly, QOL data are collected with the aim of assessing palliation. It is therefore perhaps worth spending some time looking at this aspect in detail. The simplest way of assessing whether a treatment has had an effect on symptoms is to look at the proportions of patients reporting that symptom at a certain timepoint. Such an analysis is called a landmark analysis.

Landmark Analysis

A landmark or cross-sectional analysis simply gives a snapshot of patients' QOL at a specific timepoint. One advantage is that all the data available at that timepoint can be used, but choosing the most appropriate timepoint can be vitally important as

shown by Curran et al. (2000). They compared global QOL at six different timepoints in an analysis of two regimens for locally advanced breast cancer and showed major differences in the treatments depending on the timepoint. In addition, simple landmark analyses give no indication of how patients have improved or deteriorated since the start of the trial or the start of treatment.

Improvement from Baseline

An alternative to the landmark analysis is to calculate the proportion of patients who have had an improvement in a symptom (or symptoms) by a particular timepoint. This limits the dataset to those patients with data available at both timepoints and to those who had the symptom to start with, as in a typical cancer trial, most patients will present with a multitude of symptoms and severities. Selecting a group of patients with, say, cough at baseline who complete questionnaires at the correct timepoints can drastically reduce the sample size, as shown in an extreme example in Table 2.

In an attempt to widen the definition of palliation to take into account all patients' experience, as well as duration and severity, Stephens et al. (1999) suggested the following working definition:

- Improvement—those presenting with moderate or severe symptoms must improve (i.e., the symptom must become mild or nil),
- Control—those presenting with a mild symptom must not get worse (i.e., must stay at mild or nil),
- Prevention—those presenting with no symptom must not get worse (i.e., must stay at nil).

When patients experience improvement, control, or prevention, they can be classed as successful palliation: if not, or if they die before the stated timepoint, they can be classed as failures.

Table 2. Patient sample for analysis.

	4 Drugs	2 Drugs	Total
Total Patients	154	156	310
Baseline form not completed	16	28	
Baseline form completed outside time window	18	13	
Missing data from baseline form	0	0	
No cough reported on baseline form	30	16	
Died before 3-month timepoint	39	27	
3-month form not completed	12	22	
3-month form completed outside time window	16	21	
Missing data from 3-month form	0	0	
Patients with QOL for analysis	23	29	52 (17%)

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Table 3. Response criteria for functioning scales and global QOL (adapted from Langendijk et al, Rad. Onc. 2001, 58, 257–268). Note that a low score represents poor QOL, and a high score good QOL.

Improvement
• Baseline score 0–59, with improvement of at least 5 points on at least two consecutive assessments in the first 3 months after the end of radiotherapy to a minimal value of 40.
• Baseline score 60–79, with improvement of at least 5 points on at least two consecutive assessments in the first 3 months after the end of radiotherapy.
• Baseline score 80–100, with improvement of at least 5 points on at least two consecutive assessments in the first 3 months after the end of radiotherapy.
Control
• Baseline score 60–79, with no change (i.e., <5 points) on at least two consecutive assessments in the first 3 months after the end of radiotherapy.
Prevention
• Baseline score 80–100, with no change (i.e., <5 points) on at least two consecutive assessments in the first 3 months after the end of radiotherapy.

This working definition has already been refined by another group (Langendijk et al., 2001) to include the clinically significant changes defined by Osoba et al. (1998) (see section below) as well as a definition for palliation of functioning scales and global QOL. Table 3 shows the criteria for "response," and other criteria are listed for "no response" (no change or dead without palliation), "progression," and "not evaluable."

Using these definitions, they showed that in patients with locally advanced and metastatic NSCLC (non-small cell lung cancer) radiotherapy provided excellent palliation, especially for haemoptysis, pain, and cough, and improved QOL, especially emotional and cognitive functioning.

CLINICALLY MEANINGFUL CHANGES IN QOL SCORES

Although most methods of analysis can provide us with a p value to indicate the level of statistical significance, this inevitably needs to be interpreted into clinical significance. Given a large enough sample size, even a 1% difference can be shown to be statistically significant, but such a difference would rarely be enough to change practice. There are two approaches to defining clinical significance—anchor-based (comparing QOL scores to other criteria) and distribution-based (calculating an individual patient or group effect size) (Wyrwich and Wolinsky, 2000).

Osoba et al. (1998) correlated the results from patients completing the European Organization for the Research and Treatment of Cancer Quality of Life Core Questionnaire (EORTC QLQ-C30) on repeated occasions and also rating their

perception of change since the last assessment. They found that when the functional scale scores changed by 5–10 points (on a 0–100 range), patients described their change as a little better (or worse). A change of 10–20 points correlated with a moderate change, and one of 20+ with being very much better (or worse). While King (1996), collating data from 14 studies agreed that a change of 5 or less represented a small change, she suggested that the definition for a large change differed for each scale, thus the change needed to be 16+ for global QOL, 27+ for physical functioning, but only 7+ for emotional functioning.

Such changes need to be taken into account when hypotheses are formed and sample sizes calculated.

SUMMARY

There are still no widely accepted methods of analyzing QOL data, but some general guidelines can be given:

- Whenever possible, prespecify a number of hypotheses to act as the primary and secondary QOL analyses. All other analyses should be regarded as exploratory and hypothesis-generating.
- Statistical analyses should be described in sufficient detail to allow other researchers to duplicate them.
- Use the intention-to-treat principle, and in all analyses account for all the patients.
- Consider analyzing the data in more than one way to confirm that the results were not model-dependent.
- Graphical methods may be helpful, but it is always important to specify the number of patients contributing to the plot at each key timepoint.
- Nonparametric tests such as the Wilcoxon or Mann-Whitney may be more appropriate, as QOL data are often skewed with ceiling or floor effects (e.g., patients with no symptoms cannot improve).

The importance of assessing QOL lies in presenting clinicians and patients with a full description of effects of treatments. Therefore, those of us involved in this work have a great responsibility to assess, analyze, interpret, and present such information to the best of our abilities.

REFERENCES

Bailey, A. J., Parmar, M. K., Stephens, R. J. (1998). Patient-reported short-term and long-term physical and psychologic symptoms: results of the continuous hyperfractionated accelerated radiotherapy (CHART) randomised trial in non-small cell lung cancer. *JCO* 16:3082–3093.

Burris, H., Storniolo, A. M. (1997). Assessing clinical benefit in the treatment of pancreas cancer: gemcitabine compared to 5-fluorouracil. *Eur. J. Cancer* 33 (suppl 1):S18–S22.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Cella, D. (1997). *FACIT Manual. Manual of the Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System*. Evanston, Illinois, USA: Center on Outcomes, Research and Education.

Curran, D., Aaronson, N., Standaert, B., Molenberghs, G., Therasse, P., Ramirez, A., Koopmanschap, M., Erder, H., Piccart, M. (2000). Summary measures and statistics in the analysis of quality of life data: an example from an EORTC-NCIC-SAKK locally advanced breast cancer study. *Eur. J. Ca.* 36:834–844.

De Haes, J. C. J. M., Olschewski, M., Fayers, P., Visser, M. R. M., Cull, A., Hopwood, P., Sanderman, R. (1996). *Measuring the Quality of Life of Cancer Patients with the Rotterdam Symptom Checklist (RSCL) – A Manual.* The Netherlands: Northern Centre for Healthcare Research, University of Groningen.

Fairclough, D. I., Cella, D. F. (1996). Functional assessment of cancer therapy (FACT-G): non-response to individual questions. *Qual. Life Res.* 5:321–329.

Fayers, P. M., Curran, D., Machin, D. (1998). Incomplete quality of life data in randomised trials: Missing items. *Stats. Med.* 17:679–696.

Fayers, P. M., Bleehen, N. M., Girling, D. J., Stephens, R. J. (1991). Assessment of quality of life in small cell lung cancer using a daily diary card developed by the Medical Research Council Lung Cancer Working Party. *Br. J. Ca.* 64: 299–306.

Fayers, P., Aaronson, N., Bjordal, K., Sullivan, M. (1995). *EORTC QLQ-C30 Scoring Manual*. Brussels, Belgium: EORTC Data Center.

Gelber, R. D., Goldhirsch, A. (1986). A new endpoint for the assessment of adjuvant therapy in postmenopausal women with operable breast cancer. *JCO* 4: 1772–1779, for the Ludwig Breast Cancer Study Group.

Goldhirsch, A., Gelber, R. D., Simes, J., Glasziou, P., Coates, A. S. (1989). Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. *JCO* 7:36–44, for the Ludwig Breast Cancer Study Group.

Groenvold, M., Fayers, P. M. (1998). Testing for differences in multiple quality of life dimensions: Generating hypotheses from the experience of hospital staff. *Qual. Life Res.* 7:479–486.

Hurny, C., Bernhard, J., Joss, R., Willems, Y., Cavalli, F., Kiser, J., Brunner, K. Favre, S., Alberto, P., Glaus, A., Senn, H., Schatzmann, E., Ganz, P. A., Metzger, U. (1992). Feasibility of quality of life assessment in a randomized phase III trial of small cell lung cancer. *Ann. Oncol.* 3:825–831.

Hopwood, P., Stephens, R. J., Machin, D. (1994). Approaches to the analysis of quality of life data: experiences gained from a Medical Research Council Lung Cancer Working Party palliative chemotherapy trial. *Qual. Life Res.* 3:339–352, for the MRC Lung Cancer Working Party.

Kaplan, R. M. (1993). Quality of life assessment for cost/utility studies in cancer. *Ca. Treat. Rev.* 19:85–96.

King, M. T. (1996). The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual. Life Res.* 5:555–567.

Langendijk, H., de Jong, J., Tjwa, M., Muller, M., ten Velde, G., Aaronson, N., Lamers, R., Slotman, B., Wouters, M. (2001). External irradiation versus external irradiation plus endobronchial brachytherapy in inoperable non-small cell lung cancer: A prospective randomised study. *Rad. Onc.* 58:257–268.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Machin, D., Weeden, S. (1998). Suggestions for the presentation of quality of life data from clinical trials. *Stats. Med.* 17:711–724.

Medical Research Council Lung Cancer Working Party. (1996a). Comparison of oral etoposide and standard intravenous multidrug chemotherapy for small cell lung cancer: a stopped multicentre randomised trial. *Lancet* 348:563–566.

Medical Research Council Lung Cancer Working Party (1996b). Randomised trial of four-drug versus less intensive two-drug chemotherapy in the palliative treatment of patients with small cell lung cancer (SCLC) and poor prognosis. *Br. J. Cancer* 73:406–413.

Medical Research Council Lung Cancer Working Party. (1996c). Randomised trial of palliative two-fraction versus more intensive 13-fraction radiotherapy for patients with inoperable non-small cell lung cancer and good performance status. *Clin. Onc.* 8:167–175.

Osoba, D., Rodrigues, G., Myles, J., Zee, B., Pater, J. (1998). Interpreting the significance of changes in health related quality of life scores. *JCO* 16:139–144.

Qian, W., Parmar, M. K. B., Sambrook, R. J., Fayers, P., Girling, D. J., Stephens, R. J. (2000). Analysis of messy longitudinal data from a randomised clinical trial. *Stats. Med.* 19:2657–2674.

Sadura, A., Pater, J., Osoba, D., Levine, M., Palmer, M., Bennett, K. (1992). Quality of life assessment: patient compliance with questionnaire completion. *JNCI* 84:1023–1026.

Saunders, M., Dische, S., Barrett, A., Harvey, A., Gibson, D., Parmar, M. (1997). Continuous hyperfractionated accelerated radiotherapy (CHART) versus conventional radiotherapy in non-small cell lung cancer: a randomised multicentre trial. *Lancet* 350:161–165.

Stephens, R. J., Hopwood, P., Girling, D. J. (1999). Defining and analysing symptom palliation in cancer clinical trials: a deceptively difficult exercise. *Br. J. Cancer* 79:538–544.

Stephens, R. J., Fairlamb, D., Gower, N., Maslove, L., Milroy, R., Napp, V., Peake, M. D., Rudd, R. M., Spiro, S., Thorpe, H., Waller, D. (2002). The big lung trial (BLT): determining the value of cisplatin-based chemotherapy for all patients with non-small cell lung cancer (NSCLC). Preliminary results in the supportive care setting. *Pro. Am. Soc. Clin. Onc.* 21:291a, abstract 11 on behalf of all the participants.

Wyrwich, K. W., Wolinsky, F. D. (2000). Identifying meaningful intra-individual change standards for health-related quality of life measures. *J. Eval. Clin. Practice* 6:39–49.

Zigmond, A. S., Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatr. Scand.* 67:361–370.

Received February 2003
Accepted June 2003

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016