PERGAMON

European Journal of Cancer 36 (2000) 834–844

European Journal of Cancer

www.ejconline.com

Summary measures and statistics in the analysis of quality of life data: an example from an EORTC–NCIC–SAKK locally advanced breast cancer study

D. Curran [a,*], N. Aaronson [b], B. Standaert [c], G. Molenberghs [d], P. Therasse [a], A. Ramirez [e], M. Koopmanschap [f], H. Erder [g], M. Piccart [h]

[a] *European Organization for Research and Treatment of Cancer (EORTC) Data Center, Avenue Mounier 83, Bte 11, Brussels 1200, Belgium*
[b] *The Netherlands Cancer Institute, Plesmanlaan 121, NL-1066 CX Amsterdam, The Netherlands*
[c] *Amgen, Avenue Ariane, 5, Arlanelaan, 1200 Brussels, Belgium*
[d] *Limburgs Universitair Centrum, Universitaire Campus, Building D, B3590 Diepenbeek, Belgium*
[e] *ICRF Psychocial Oncology Group, St Thomas Hospital, London SE1 7EH, UK*
[f] *IMTA, Erasmus University, Rotterdam, The Netherlands*
[g] *Amgen, Thousand Oaks, CA, USA*
[h] *Institut Jules Bordet, Rue Heger-Bordet, 1, Brussels 1000, Belgium*

Received 28 May 1999; received in revised form 10 February 2000; accepted 29 February 2000

Abstract

Quality of Life (QL) is now included as an endpoint in many phase III cancer clinical trials. Numerous statistical techniques have been presented in the literature to analyse QL data but there is still no agreement as to what is the optimal approach of analysis. In this paper we, therefore, present and compare various techniques which have all appeared in the literature and which may be globally described as summary measures and summary statistics. These techniques are illustrated using data from an EORTC clinical trial in locally advanced breast cancer (EORTC trial 10921). It is also explained in this paper how and when these techniques may be used in other cancer settings. For EORTC trial 10921, it is shown that by choosing different techniques different conclusions may be drawn concerning the QL outcome. This highlights the importance of choosing an appropriate primary statistical method and for describing it a priori in the protocol and analysis plan. In this paper, we show the importance of performing sensitivity or supportive analysis to support conclusions drawn from the primary analysis. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Quality of life; EORTC QLQ-C30; Summary measures; Summary statistics

1. Introduction

During the last two decades, numerous instruments have been developed to assess patients' quality of life (QL). In the intervening years assessment of QL has rapidly become an integral part of clinical research resulting in many studies yielding vast quantities of data. However, the question as to what is the best way to analyse and present the results has not been sufficiently addressed. Researchers have sought a practical solution to the conflict of complexity of QL datasets and the desire to simplify the presentation of results. Nevertheless, controversies surrounding quality of life analyses have remained, mainly due to the fact that a standard questionnaire consists of numerous categorical scales, assessed at frequent time points during the study, and also because patients may drop out of the study at various times.

It is widely accepted that QL is composed of various dimensions. However, to simplify analyses and to ease interpretation one scale may be chosen as the primary outcome. The other scales are then considered secondary in nature and analysed in an exploratory fashion. For example, when the EORTC Quality of Life Questionnaire Core 30 (QLQ-C30) is included in a clinical trial, the global health/QL scale may be taken as the

* Corresponding author. Tel.: +32-2-774-1670; fax: +32-2-772-6701.
E-mail address: dcu@eortc.be (D. Curran).

PII: S0959-8049(00)00056-3

primary QL endpoint. If a difference is observed in this scale between the treatment groups, then it is important to attempt to explain why the difference occurs. This can be done by analysing the remaining scales, such as physical, emotional and role functioning in an exploratory way. Thus the primary analysis may be seen as confirmatory where the question "Is QL different between the two groups?" is addressed, whereas in the exploratory data analysis the objective is to support the initial finding and to generate research hypotheses rather than draw definitive conclusions from the results.

Important questions are related to the repeated structure of the data over time. Summary measures have been widely accepted as useful methods for reporting results from longitudinal studies [1,2]. In essence a summary measure collapses the complete set of measurements of an individual into a single number. A summary measure should be chosen to reflect some important aspect of the repeated measurements. For example, in oncological clinical trials, data on toxicity are often summarised by taking the worst value recorded for each patient during the entire treatment period.

Within clinical trials, QL is usually reported by each patient on a self-report questionnaire at repeated timepoints before, during and after treatment. Summary measures may be useful for simplifying the repeated structure of the data. A few such measures that could be considered are the mean, median and the minimum or maximum score recorded for each individual patient. The summary measures for all patients are then analysed using an appropriate univariate method.

Tannock and colleagues [3] in a clinical trial of prostate cancer patients, used two summary measures in analysing the QL data. Each of the patient's scores for each QL domain was summarised using the median and the best score. These were subsequently converted to median and best change scores by subtracting the patient's baseline score. Differences in the summary scores between the two treatment groups were assessed with the Wilcoxon rank-sum test.

Hollen and associates [4] used the Area Under the Curve (AUC) method as a summary measure. The AUC may be interpreted as a weighted average of the scores at each assessment time point. It is frequently used in other fields of research such as pharmacology, economics and engineering. Usually, the AUC is calculated only for the period of time during which measurements are available for the patient. However, Hollen and associates [4] extended this method to also take into account patients who dropped out of the study by imputing a score of zero for those patients after dropout.

A distinction should be made between summary measures and summary statistics. A summary measure reduces the measurements for one individual to one single number, whereas a summary statistic reduces the measurements of a group of individuals to one number. Similarly, a summary statistic may be a summary of the group differences in QL between two treatment strategies. For example, several authors have compared treatments with respect to QL at individual time points (e.g. using a Student *t*-test or a Wilcoxon test). Seymour and coworkers [5], in a study of colorectal cancer patients, presented summary statistics at each time point for each treatment group and used exact Chi-squared tests to compare the QL scores in the two treatment groups. Although the sample size may vary at each time point, this method makes use of all available data. However, there are problems with using this method, which are described later in the paper.

Using a practical example from an EORTC clinical trial, this paper investigates a number of methods of analysis that have all been presented in the literature. A number of summary measures and summary statistics are discussed and it is shown how the choice of method of analysis influences the study results. Examples are provided where it may be useful to include summary statistics and measures to reflect an important aspect of the study. The advantages and disadvantages of each method and the basic assumptions that are required when using these methods are discussed.

2. Patents and methods: dataset

EORTC trial 10921 is an international, intergroup dose-intensified study [6]. This randomised phase III study was designed to compare six 4-weekly cycles of CEF (cyclophosphamide, epirubicin and 5-fluorouracil (5-FU) versus six 2-weekly cycles of dose-intensified EC (epirubicin and cyclophosphamide) + G-CSF (filgrastim) in patients with locally advanced/inflammatory breast cancer. Thus, the expected duration of standard treatment is 24 weeks compared with 12 weeks in the intensified treatment arm. Between June 1993 and April 1996, 448 patients were entered in the trial with 224 patients being randomised into the CEF arm and 224 into the EC + G-CSF arm. The main endpoint of the trial was progression-free survival. To date, duration of survival and progression-free survival are not significantly different between the two treatment arms [6].

QL was considered to be a mandatory part of the protocol. The QL assessment consisted of two generic questionnaires, the EuroQoL [7] and the RAND MOS [8], a cancer-specific questionnaire, the EORTC QLQ-C30 [9] and a study-specific breast module. The global health status/QL scale from the EORTC QLQ-C30 was specified as the QL domain on which the main analysis would be performed. This scale was constructed using the scoring procedures for the EORTC Core Quality of Life Questionnaire EORTC QLQ-C30 version 1.0 [10], i.e. the scale score was calculated by averaging items within the scale and transforming the average score linearly to a 0–100 scale, with higher scores representing a better global health status/QL.

The planned schedule of assessment in both treatment arms was as follows: at randomisation, every month for the first 3 months, every 3 months for the first year, at 18 months and every 8 months thereafter until disease progression. The current analysis concentrates on the QL assessments during the first year. A window (time-frame) for acceptance of questionnaires was defined for each assessment. To allow for some delay in the schedule of treatment (and hence QL assessment) more time was allowed after than before the scheduled assessment point (e.g. at 4 weeks questionnaires were accepted if they were completed within one week before and up to 2 weeks after the scheduled assessment). This reduces the possibility of patients being omitted from the analysis because of delayed chemotherapy cycles.

3. Results

Of the 448 patients included in the trial 11 patients were considered ineligible (8 due to disease stage, 1 due to prior treatment for breast cancer and 2 due to previous or concurrent malignancy). For 2 patients eligibility was not verifiable due to inadequate source documentation. 17 patients from two institutions were excluded from the QL study, as officially translated EORTC QLQ-C30 questionnaires were not available for these countries during the study. A further 15 patients were excluded because they were judged unfit to complete the QL questionnaires (9 in the CEF arm and 6 in the EC + G-CSF arm). A total of 403 patients were eligible for the QL analysis, 199 in the CEF arm and 204 in the EC + G-CSF arm.

The compliance with the QL assessments during the first 12 months of the study is presented in Table 1. Some patients completed more than one valid QL questionnaire within a given time window. For these patients the questionnaire which was completed first during this period was retained. The main reason for patients going off study was due to progression [6]. Table 2 presents the patterns of completed questionnaires. 93 patients completed QL questionnaires at all seven assessment time points. Monotone dropout patterns (i.e. a complete series of questionnaires before dropout) were observed in 189 cases (this includes the latter 93 patients). Intermittent missing questionnaires was also a problem with 119 patients having exactly 1 missing questionnaire and the remaining 90 patients having more than 1 missing questionnaire in a series before dropout. 5 patients did not complete any questionnaires during this period.

In 1996 the steering committee for EORTC trial 10921 drew up an analysis plan for the study. It was decided that the primary analysis would be based on an

Table 1
Compliance with QL assessment by treatment arm

Months	0	1	2	3	6	9	12
CEF							
Expected	199	199	199	195	179	167	155
Received	169	157	148	156	124	114	103
%	85	79	74	80	69	68	66
EC + G-CSF							
Expected	204	204	202	199	194	185	170
Received	173	169	158	141	133	127	104
%	85	83	78	71	69	69	61

Table 2
Patterns of missing data

Months								Months							
0	1	2	3	6	9	12	Frequency	0	1	2	3	6	9	12	Frequency
*	*	*	*	*	*	*	93	*	*				*		2
*	*	*	*	*	*		27	*	*			*	*		2
*	*	*	*	*			21	*	*		*			*	2
*	*	*	*				16	*		*					2
*	*	*					14	*		*		*			2
*	*						8	*		*	*		*		2
*							10	*		*	*	*			2
*	*	*	*	*		*	14		*	*	*		*	*	2
*	*	*	*		*	*	11		*		*				2
*		*	*	*	*	*	11		*		*	*	*		2
*	*		*				11			*			*	*	2
*	*		*	*	*		8			*	*				2
*	*	*		*	*	*	7					*		*	2
	*	*	*	*	*	*	7		*	*		*	*	*	1
*	*	*		*			7	*	*		*	*		*	1
*	*		*	*	*	*	6	*	*			*			1
*	*	*	*		*		6	*		*				*	1
*	*		*	*			6	*		*		*	*		1
	*	*					5	*				*			1
*	*	*	*			*	4	*			*				1
*			*	*	*	*	4	*			*	*			1
*		*	*				4		*	*				*	1
*			*		*	*	4		*	*		*			1
	*	*	*		*		4		*	*		*		*	1
*	*	*			*	*	3		*	*	*				1
*	*		*		*	*	3		*	*	*	*		*	1
*		*	*	*	*		3		*						1
*	*				*	*	3		*				*		1
*		*			*	*	3		*				*	*	1
	*	*	*	*			3		*			*	*		1
*	*			*	*	*	2		*			*	*	*	1
		*	*	*	*	*	2			*					1
*	*	*		*	*		2			*				*	1
*		*	*		*	*	2			*		*	*	*	1
	*	*	*	*	*		2			*	*		*	*	1
	*		*	*	*	*	2			*	*	*			1
*		*		*	*	*	2			*	*	*		*	1
*	*	*		*		*	2					*	*	*	1
*	*	*			*		2				*	*			1
*	*					*	2								

AUC analysis. This was mainly due to the expectation that the intensified treatment would initially result in a reduced QL but patients would recover more rapidly due to the shorter duration of treatment, whereas patients in the standard arm would experience side-effects of treatment over a longer period of time. Therefore, it was assumed that the most appropriate method of balancing the short-term side-effects of intensified treatment (EC+GCSF) with the extended side-effects of the standard treatment (CEF) was to perform an AUC analysis. In this paper we respect the original analysis plan whilst examining other methods of analysis which have all been presented in the literature.

3.1. Summary measures

Several types of summary measures may be employed in the analysis of QL data. These can be categorised as follows: (1) simple summary measures, e.g. minimum, maximum, median or mean QL score for a patient; (2) time to occurrence of event where, for example, the time to observation of the first minimum or maximum score is taken; (3) area under the curve where both time and QL are summarised into one single number for each individual.

3.1.1. Simple summary measures

As mentioned, examples of simple summary measures are the minimum, maximum, median and mean QL score for a patient. In oncological clinical trials where a new experimental chemotherapy is being investigated one may wish to investigate if the experimental treatment is less toxic than the standard treatment whilst achieving equivalent efficacy results. In such trials where there is an interest in reducing toxicity and maintaining an acceptable QL, the worst symptom score may be of particular importance as a summary measure.

In contrast, in clinical trials involving patients with advanced disease (e.g. symptomatic disease), the treatment provided may be palliative in nature, i.e. directed at symptom relief. In such trials, where the primary objective may be to reduce the patient's suffering and thus improve the quality of their remaining life, a useful summary statistic could be the best score with respect to symptom relief or the best QL score.

In EORTC trial 10921, a trial designed to show superiority, a plot of the individual patient scores (data not shown) indicated a good deal of variation in within-patient scores in both treatment arms. Therefore, it was thought that the mean or median score over all sequences would provide useful insight into the patients' QL. The mean score also gives an indication of the frequency of episodes or intensity of problem over time. For example, two treatment groups may have comparable best or worst scores but one treatment group may have consistently lower scores. Since the mean is an average of all observed scores it would reflect this phenomenon. In contrast, when two means are being compared it is also important to investigate the spread of scores (i.e. variance). If two treatments provide similar means but the variance is significantly larger in one treatment group, from a conservative point of view, one might prefer the treatment which provides more consistent results as it minimises worst case scenarios.

Fig. 1 presents these summary measures for patients in trial 10921 during the first year. The Wilcoxon rank-sum test was used to compare QL scores in the two treatment groups. A significant group difference was observed in terms of minimum ($P<0.001$), mean ($P=0.016$) and median QL scores ($P=0.041$) during the first year in favour of the CEF treatment arm. Note, that the differences in mean QL score and median QL score, although statistically significant, were relatively small. The null hypothesis of no treatment difference could not be rejected for the maximum summary measure. These results would suggest that there is a dip in the QL score during the first year in the EC+G-CSF arm.

3.1.2. Time to occurrence of a summary measure

Time to event analyses are frequently used in cancer clinical trials. In most cases the event is death or disease progression and the associated time periods are referred to as duration of survival and time to disease progression, respectively. Some trials have also investigated time to a certain increase in a tumour marker such as prostate specific antigen (PSA) in prostate cancer. In QL research it may also be useful to use this approach to investigate the time at which QL is at its worst or at its best, or when a certain decline in QL scores is observed.

In the analysis plan of trial 10921, it was hypothesised that QL would initially deteriorate due to treatment toxicity, in both treatment arms, but that it would increase thereafter due to relief of symptoms related to the tumour. One might expect that this increase would occur more rapidly in the intensified treatment arm due to the shorter duration of treatment and the fact that treatment included G-CSF. Thus, an interesting question was to investigate at what point in time during the first year did patients report their maximum QL score. Towards this end, the maximum QL score during the first year was obtained for each patient. An event was defined as a maximum QL score greater than the patient's baseline QL score. If the patient's maximum score was not greater than that at baseline or if the patient dropped out before observing a maximum score greater than that at baseline the patient was censored at the time of the last available assessment during the first year. The time to event was defined as the time to the first maximum QL score. In Fig. 2, maximum QL scores tended to be observed earlier in the CEF arm. However, at months 6 and 9, there was a greater tendency for maximums to be reached in the EC + G-CSF arm. No significant difference was observed between the two treatment groups ($P=0.507$). Approximately 50% of patients in both treatment arms observed a maximum score greater than baseline during the first year.

In the above approach, due to the original categorical nature of the EORTC global health status/QL scale, a score greater than baseline can be interpreted as an improvement of at least 8 points on a 0–100 scale. A similar approach would be to define a specific minimum level of improvement in QL (e.g. a change of 20 points or 0.5 of a standard deviation) and to define the time to event as the time to reach such a minimal improvement.

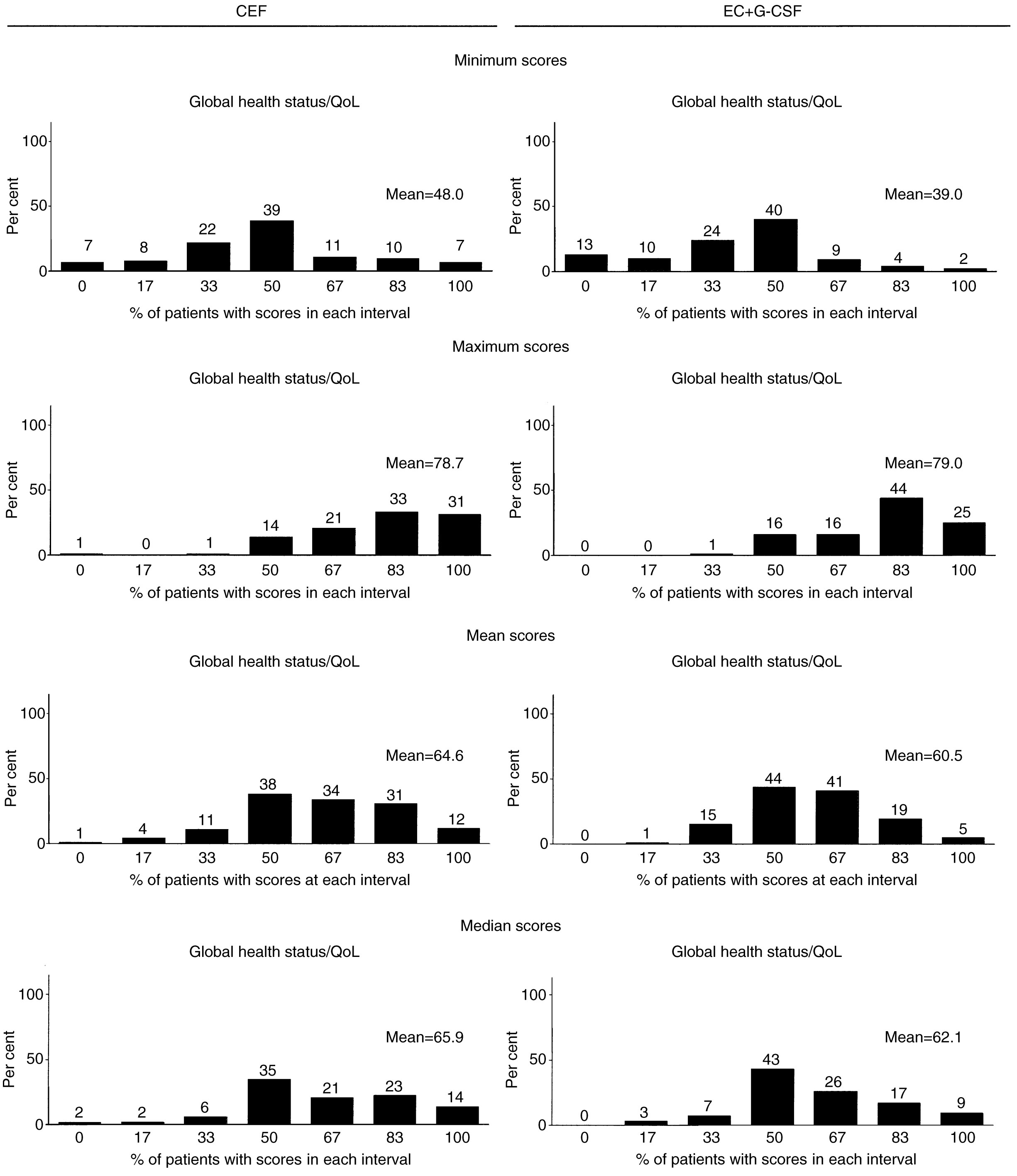

Fig. 1. Summary measures for the global health status/QL score. Note: for presentation purposes the scores have been grouped into equally spaced intervals with midpoints 0, 17, 33, 50, 67, 83, 100. The x axis represents the percentage of patients with scores in each interval.

3.1.3. Area under the curve

The area under the curve is calculated by summing areas under the graph between each pair of consecutive observations. Thus, the AUC is a weighted average of the QL scores at each individual time point weighted by the time between observations (see Appendix A). In QL analysis, time may be considered as discrete or continuous, i.e. the time of assessment may be taken as the planned time of assessment (e.g. time points 0, 1, 2 months which is discrete) or as the actual observed assessment time points (e.g. −1, 29, 61 days which is continuous). In trial 10921 we chose to treat time as a continuous variable. The AUC was compared between the two treatment groups using the Wilcoxon rank sum test. No significant difference was observed between the two groups ($P=0.882$).

One of the advantages of the AUC method is that a sensitivity analysis may easily be performed to investigate the change in QL between two consecutive assessments. For example, in trial 10921, QL was assessed at months 1, 2 and 3 and then at months 6, 9 and 12. Of particular importance was the question "Did patients recover rapidly after treatment with EC + G-CSF?". Generally, the AUC is calculated assuming a linear change in QL scores between consecutive assessment time points. However, the formula for calculating the AUC may be changed to allow the rate of change between assessments to occur non-linearly (see Appendix A and Fig. 3). A sensitivity analysis may be performed to investigate if allowing the rate of change between assessments to vary results in different conclusions with respect to treatment effect. An alternative sensitivity analysis would be to investigate what is the effect of dropout, analogous in principle, to the approach taken by Hollen and colleagues [4] who imputed a score of 0 on the day of death.

Note, the AUC method should give approximately the same results as taking the mean as a summary measure if the time between assessments is equal. When using the mean each score is given equal weight whereas the AUC method weights the scores according to the time between assessments. We used the trapezium rule to calculate the AUC and compared the resulting summary measures using the Wilcoxon rank sum test. No significant difference was observed between the two treatment arms. Recall that when the mean was used as a summary measure there was a significant difference between the two treatment groups ($P=0.016$). When using the AUC, the area under the curve between baseline and month 3, which contains four assessments, is given the same weight as the AUC between months 6 and 9, which contains two assessments. Thus, the area under the curve provides a more balanced estimate of the overall QL during the first year than does the mean.

3.1.4. Limitations of summary measures

When using summary measures, as with any type of statistical analysis, care has to be taken that bias is not being introduced. For example, where the worst score for a symptom is taken as a summary measure, the results may be biased if patients with a high level of a symptom are unable to complete a self-assessment questionnaire and are thus not able to report their worst level of symptoms. This would lead to a biased estimate of the level of symptoms in each treatment arm. Similarly, this may result in a biased estimate of the relative effect of one treatment group versus the other.

The schedule of assessments should be selected to optimise information regarding QL in the two treatment arms. Likewise, the choice of summary measure is closely related to the schedule of assessments and should be chosen to reflect an important feature of the patients' QL. In trial 10921, QL assessments at months 4 and 5 might have been warranted to obtain a clearer picture with respect to how rapidly patients recovered from intensive treatment.

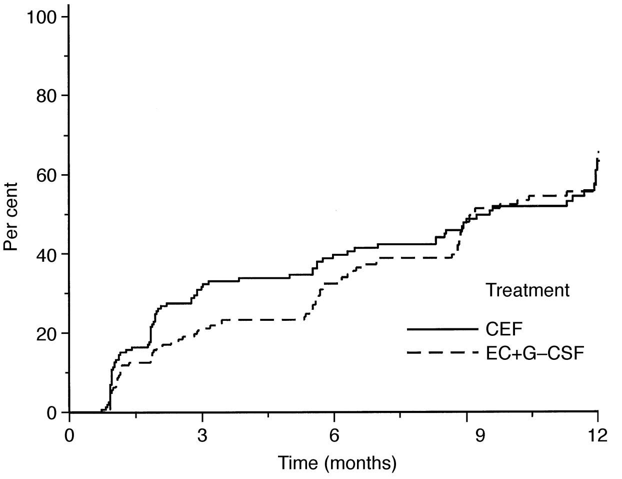

Fig. 2. Time to maximum QL score greater than baseline.

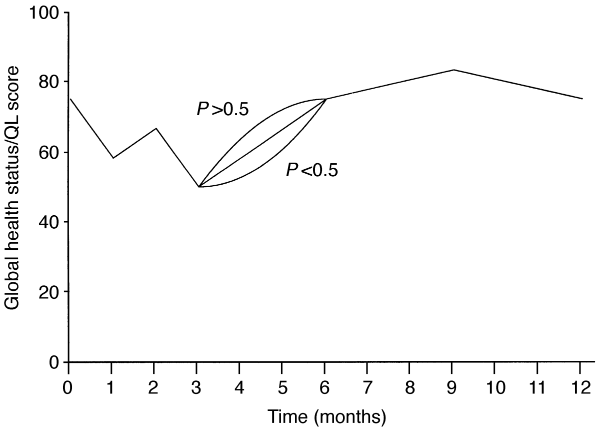

Fig. 3. Global health status/QL score during the first year for an individual patient.

When using summary measures, bias may be introduced into the comparison if the follow-up periods are not similar in the two treatment arms. Additionally, the rates of completing questionnaires should be high and equivalent across treatment arms. In trial 10921 the progression-free survival and QL compliance were similar in the two treatment arms. Summary measures may not be appropriate in studies where dropout of patients is high as they ignore the problem of incomplete data.

In "time to occurrence of event" analyses there maybe some difficulty in defining an event and defining censoring. In the above analysis, it was assumed that the majority of patients would observe the event of interest during the first year and thus censoring would have less impact on the treatment comparison. However, if the timing of censored observations is different between the two groups there may be problems with the interpretation of results.

Generally, summary measures such as the minimum and maximum score are very sensitive to outliers (i.e. extreme observations). When analysing categorical data this may not be a problem. However, for continuous data one might consider using more robust estimators (e.g. mean or median).

3.2. *Summary statistics*

Since QL measurements are typically obtained via repeated assessments over time, it is generally assumed that QL data should be analysed as such, taking the repeated measurements into account. However, this is often hampered by the structure of the data; i.e. QL data are usually measured on ordered categorical response scales and a proportion of questionnaires will be missing both intermittently and due to dropout of patients from the study. Statistical techniques for repeated, ordered categorical, incomplete data are limited. Many statistical analysis strategies assume that complete balanced data matrices are available and that data are normally distributed. This has led QL researchers to perform separate analyses at each assessment time point. This method of analysis is usually referred to as cross-sectional analysis or available case analysis as it uses all available data at each assessment time point.

3.2.1. *Cross-sectional analysis*

If the distribution of QL scores is approximately normally distributed, it may be appropriate to perform simple Student *t*-tests within each cross-sectional analysis. Often non-parametric tests, such as Wilcoxon or Mann–Whitney tests, may be more appropriate, in that many QL questionnaires yield skewed distributions with notable ceiling or floor effects (i.e. the proportion of patients with either none or severe problems). If there is a large difference in the mean QL score between the two treatment groups one might also expect that the variance may be quite different in both treatment groups, in particular with skewed distributions, suggesting that if a standard Student *t*-test is to be performed the variance in each group should first be investigated.

In trial 10921, we compared the two treatments with respect to global health status/QL score at each individual time point using a Wilcoxon test. The results are presented graphically in Fig. 4. The plot indicates that, compared with the standard regimen, the intensified regimen had a significant negative impact on QL during the first 3 months. At month 6 the QL score returned to pretreatment levels in the intensified arm, whilst patients in the standard arm tended to have a poorer QL score. No significant differences were observed between the two groups at months 9 and 12.

The main disadvantage of this method is that different sets of patients contribute at different time points depending on the pattern of missing data. Thus, this method yields problems of comparability across time points. Additionally, it does not control for any potential biases in the treatment comparisons, which may occur due to dropout of patients. The procedure of examining differences between groups of patients at each assessment time point also leads to inflated Type I and II errors due to multiple testing. An adequate adjustment of the significance level of each test [11,12], or a combination of individual test statistics into a global statistic is then essential [13]. If many statistical tests are being performed, it is possible to use a more restrictive significance value such as $P<0.01$ or 0.001, thereby reducing the risk of Type I errors. In the next section we will discuss the Wei–Johnson procedure which allows the per time point test statistics to be combined into one overall test statistic [13].

3.2.2. *Wei–Johnson*

Overall tests of significance are generally preferable to comparisons per time point. Overall tests allow general

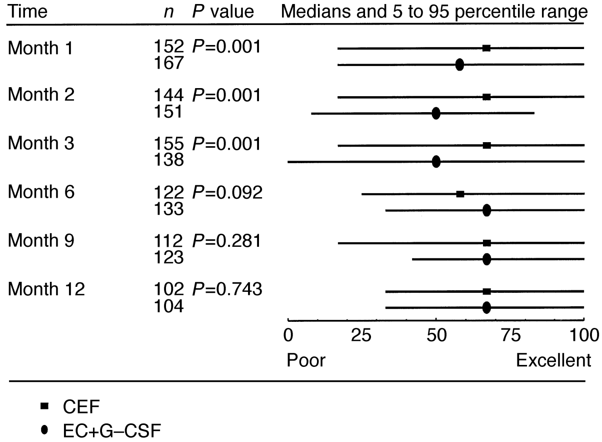

Fig. 4. Cross-sectional analysis of global health status/QL score during the first year.

statements about effects, are statistically more powerful and provide a safeguard against multiple comparisons. When overall tests yield statistically significant results, they can be followed by exploratory comparisons per time point. Wei and Johnson proposed a test [13], which allows cross-sectional tests to be combined in an overall treatment comparison. They illustrated how cross-sectional Wilcoxon tests, Student *t*-tests or tests for 2×2 tables could be combined. In trial 10921, cross-sectional analysis to compare the QL scores between the two treatments at each of the time points were performed using the Mann–Whitney test. These were then combined using the Wei–Johnson method as illustrated in Appendix B. The estimate of the Wei–Johnson test statistic, when taking equal weights for each time point (i.e. $w_j = 1$ for all j), is $W = 3.404$ ($P < 0.001$). Defining the weights relative to the time period between measurements for the six assessments at 1, 2, 3, 6, 9 and 12, respectively yields $w = \{0.5\ 0.5\ 0.5\ 1.5\ 1.5\ 1.5\}$. Note the average of all the weights (w_j's) should be 1 (since the first three assessments are taken 1 month apart they receive a weight of 0.5 and the last 3, taken 3 months apart, receive a weight of 1.5). In this setting, the Wei–Johnson test statistic is $W = 1.812$ ($P < 0.070$). Various alternative techniques for estimating weights could be employed including estimating weights based on the number of patients contributing to the analysis at each time point.

Note, the Wei–Johnson procedure has all the disadvantages of the cross-sectional analysis presented in Section 3.2.1 except that it produces an overall test and therefore reduces the number of statistical comparisons. Although the Wei–Johnson procedure allows one to analyse the data in a longitudinal fashion it does not tell us anything about the correlation structure. For example, one might expect that QL measurements taken close together are strongly correlated whilst measurements taken at larger intervals are less correlated. Whilst, cross-sectional studies are useful in describing between-patient variability, explaining within-patient variability necessitates the study of the repeated assessments over time [14].

3.3. Categorical data

Although all of the above methods were illustrated using the global health status/QL scale, most of the methods (perhaps with the exception of the AUC method) can be applied to the other scales which have fewer potential categories such as the single item scales of the EORTC QLQ-C30. For single-item scales with binary or ordered categorical response scales another summary measure could be the frequency of events. This could be particularly useful for scales assessing symptoms or toxicity. For example, in the QLQ-C30 one might be interested in the number of times patients responded that they had "Quite a bit" or "Very much" trouble sleeping during the treatment period.

When performing cross-sectional analyses with ordinal response categories an approach that is useful is to compare proportions of patients with a certain category. For example, instead of comparing the distribution of insomnia scores between the two treatment groups, one could calculate the proportion of patients in each group who report having "Quite a bit" or "Very much" trouble sleeping. A proportion is one summary statistic that is easy to describe and facilitates understanding of results. The Wei–Johnson procedure may also be used to combine tests of proportions at several time points into one overall test. Additionally, two proportions can easily be compared using a Chi-square test. Cumulative proportions are of particular relevance when dealing with ordinal data and may be analysed using odds ratios.

4. Discussion

Statistical methodology for analysing QL data has developed rapidly over the last few years. In fact, the analyses may be performed in several ways based on different assumptions. However, due to the complicated nature of QL datasets, care should be taken when choosing a method for analysis since there is a risk of drawing incorrect conclusions if inappropriate statistics are used. Studies with inappropriate statistical methodology, due to lack of expertise, may be scientifically useless and hence an unethical waste of valuable resources. Additionally, erroneous conclusions may lead to inappropriate treatment of subsequent patients.

In this paper, we investigated a number of different summary measures and statistics for QL data. We illustrated that the conclusions drawn may be different depending on the type of analysis performed. In our analysis plan the primary analysis was defined as the AUC method. Using this approach, no significant overall difference was observed between the two treatment arms during the first year after randomisation. By definition, summary measures and statistics do not use all the data collected and as such they may be considered wasteful. For example, they do not take into account how patients' scores change over time. Due to this, study conclusions should not be based solely on one summary measure or summary statistic, but should be supported by additional analyses in the form of a sensitivity analysis.

The additional analyses presented in this paper were useful in gaining insight into the data and in understanding the impact of treatment on the patients' QL. Significant differences, at the 5% level, were observed in favour of the CEF arm when the minimum, mean and median summary measures were used. No significant differences were observed when the maximum score, time to maximum score and AUC methods were employed. When using summary statistics based on cross-sectional analysis significant differences were observed at months 1, 2 and 3 in favour of the CEF arm. However no significant differences were observed at later time points. When the Wei–Johnson test statistic test, with equal weights, was used to combine cross-sectional results into one single summary measure, a significant difference in favour of CEF was found.

In conclusion, the main differences between the two treatment groups occurs during the first 3 months where QL scores are significantly lower in the EC + G-CSF arm (illustrated by the cross-sectional analysis and minimum scores). The scores are borderline significantly different at 6 months in favour of EC + GCSF, which is probably due to the fact that patients were still receiving treatment in the CEF arm whereas patients in the EC + G-CSF arm had recovered from treatment toxicities. At later time points, there are no significant differences between the two groups. As the summary measures: minimum, maximum, mean and median attach equal weight to each assessment they ignore the time between assessments. The Wei–Johnson method allows one to explore how associating different weights to the various assessments yields different conclusions. The AUC method yields an overall score for each patient, which is calculated as a cumulative weight of all QL scores weighted according to the time between assessments. Thus, the difference between methods can be explained in part by the weighting of time, e.g. short-term differences versus overall differences. Unfortunately, the QL study design did not request QL assessments at months 4 and 5 during which patients in the CEF arm continued to receive treatment. One could hypothesise that patients in the intensified treatment arm recovered rapidly from treatment toxicities after going off-treatment due to the fact that the intensified treatment included G-CSF.

It is clear that depending on the method of analysis different conclusions could be reached. This highlights the importance of choosing appropriate statistical techniques. Secondly, as many statistical techniques may be employed it is important to describe the primary statistical analysis a priori in the study protocol. In the statistical analysis plan for this study, we specified that the primary analysis would be based on the AUC method. In retrospect, for study 10921, the authors concur that the AUC method provides a valid overall assessment of QL during the first year. In general, the choice of summary measure will depend on the patient population (e.g. lung or breast cancer), the treatment (curative or palliative) and the design of the clinical trial for the clinical outcome (equivalence or difference).

As was mentioned earlier, summary measures may be useful to reflect an important aspect of the data and they have the advantage of being quite easy to calculate. Once the appropriate summary measure has been obtained for each patient the values can be used in a simple treatment comparison. Similarly, summary statistics are easily obtained for each assessment time point. In general, summary measures and summary statistics assume that there is no bias in the treatment comparisons due to intermittent missing data or due to patients dropping out of the study. This may not be the case if patients do not complete the questionnaire because they are unfit to do so or if they drop out of the study due to progressive disease or worsening clinical condition. In practice it is usually impossible to conclude definitively whether dropout causes a bias in the treatment comparison or not, since the required information is not available. It is therefore important to prospectively collect the reasons for missing QL assessments. Curran and associates [15] discussed how the bias due to dropout may be investigated and how it may affect the study results [16]. They concluded that the main issue is whether there is differential dropout in the two groups. This is likely to occur when the clinical outcomes (e.g. time to progression or progression-free survival) differ between the two treatment groups.

Recently, much attention has been given to longitudinal (repeated measures) analyses with possibly intermittent missing data and missing data due to dropout of individuals from a study [14,17,18]. This includes graphical techniques to explore the longitudinal structure of the repeated measurements followed by statistical modelling of the longitudinal measurements using selection models and pattern-mixture models [19]. Unlike summary measures and summary statistics these approaches make maximum use of the available data. For example, they allow one to examine the correlation structure between repeated assessments during model fitting and to describe the between-patient variability and within-patient variability; they take into account the fact that patients with poorer scores may be more likely to drop out earlier and therefore they produce potentially less biased results; they allow the dropout rate to be different between the two treatment arms reducing the bias caused by dropout; they do not need to assume linear change of QL scores over time; and they can circumvent the problems which arise when the treatment schedule, and thus the QL assessment schedule, is not synchronised for treatment arms. Thus, using these more sophisticated techniques provides a clearer overall picture of the impact of disease and treatment on the QL of patients. In the field of QL research, longitudinal analysis of both categorical and continuous measurements is a rapidly developing area with many methods currently being explored [19–24]. Although, these techniques require a higher level of statistical sophistication from the analyst the results can easily be disseminated and interpreted by a non-statistical audience.

Acknowledgements

The EORTC Data Center is grateful to Amgen and Pharmacia Upjohn who provided support for this study.

Appendix A

Using the trapezium rule the area under the curve is approximated as follows:

$$AUC = \frac{1}{2}\sum_{j=0}^{n-1}\left(t_{j+1} - t_j\right)\left(y_j + y_{j+1}\right) \quad (1)$$

where y_j is the scale score at time t_j. In the example provided in Fig. 3, measurements are available at one monthly intervals for the first 3 months (thus $t_{j+1}-t_j=1$) and at 3-monthly intervals during the subsequent periods (thus $t_{j+1}-t_j=3$). For an individual patient with global health status/QL scores: 75 58 67 50 75 83 75, respectively at the seven assessment time points the AUC is calculated as follows:

$$\begin{aligned}AUC &= \frac{1}{2}\sum_{j=0}^{n-1}\left(t_{j+1} - t_j\right)\left(y_j + y_{j+1}\right)\\ &= \frac{1}{2}\big((75+58)+(58+67)+(67+50)\\ &\quad + 3\times((50+75)+(75+83)+(83+75))\big)\\ &= 849\end{aligned}$$

However, the formula in Eq. (1) may be changed to incorporate a parameter p ($p \in (0,1)$), which allows the rate of change between assessments to occur non-linearly (see Fig. 3).

$$\begin{aligned}AUC = \sum_{j=0}^{n-1}&\left(t_{j+1} - t_j\right)\\ &\times\left(\min\left(y_j, y_{j+1}\right) + p\left(abs\left(y_j - y_{j+1}\right)\right)\right) \quad (2)\end{aligned}$$

For the special case where $p=1/2$, equations (1) and (2) are equivalent. A sensitivity analysis may be performed to investigate if changing the parameter p modifies the conclusions with respect to treatment effect. Where *abs*, absolute value; min, minimum.

Appendix B

Let U_j be the Mann–Whitney test statistic obtained for each cross-sectional analysis j, then the Wei–Johnson statistic is defined as

$$W = \left(\sum_j w_j U_j\right) / (w'Vw)^{1/2}$$

where V is a covariance matrix of $U=(U_1,U_2,.....,U_j)$ and w is a vector of weights (w_j) given to each cross-sectional analysis j.

References

1. Fairclough DL. Summary measures and statistics for comparison of quality of life in a clinical trial of cancer therapy. *Stat Med* 1997, **15**, 1197–1209.
2. Matthews JN. A refinement to the analysis of serial data using summary measures. *Stat Med* 1993, **15**, 27–37.
3. Tannock IF, Osoba D, Stockler MR, *et al.* Chemotherapy with mitoxantrane plus prednisone alone for symptomatic hormone-resistant prostate cancer: a Canadian randomized trial with palliative endpoints. *J Clin Oncol* 1996, **14**, 1756–1764.
4. Hollen PJ, Gralla RJ, Cox C, Eberly SW, Kris M. A dilemma in analysis: issues in serial measurement of quality of life in patients with advanced lung cancer. *Lung Cancer* 1997, **18**, 119–136.
5. Seymour MT, Slevin ML, Kerr DJ, *et al.* Randomized trial assessing the addition of interferon alpha 2a to fluorouracil and leucovorin in advanced colorectal cancer. Colorectal Cancer Working Party of the United Kingdom Medical Research Council. *J Clin Oncol* 1996, **14**, 2280–2288.
6. Therasse P, Mauriac L, Welnicka M, *et al.* Neo-adjuvant dose intensive chemotherapy in locally advanced breast cancer (LABC): An EORTC–NCIC–SAKK randomized phase III study comparing FEC (5-FU, epirubicin, cyclophosphamide) vs high dose intensity EC+G-CSF (Filgrastim). *J Clin Oncol* 1998, **17**, 124.
7. The Euroqol Group. Euroqol — a facility for the measurement of health related quality of life. *Health Policy* 1990, **16**, 199–228.
8. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992, **30**, 473–483.
9. Aaronson NK, Ahmedzai S, Bergman B, *et al.* The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993, **85**, 365–376.
10. Fayers PM, Aaronson NK, Bjordal K, Curran D, Groenvold M. *EORTC QLQ-C30 Scoring Manual.* 2nd edn. Brussels, EORTC, 1995.
11. Pocock SJ, Geller NI, Tsiatis AA. The analysis of multiple endpoints on clinical trials. *Biometrics* 1987, **43**, 487–498.
12. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrica* 1988, **75**, 800–802.
13. Wei LJ, Johnson WE. Combining different tests with incomplete repeated measurements. *Biometrics* 1985, **72**, 359–364.
14. Diggle PJ, Liang KY, Zeger SL. *Analysis of Longitudinal Data.* Oxford, Clarendon Press, 1994.
15. Curran D, Molenberghs G, Fayers P, Machin D. Aspects of incomplete quality of life data in randomized trials: II: missing forms. *Stat Med* 1998, **17**, 697–709.
16. Curran D, Bacchi M, Hsu Schmitz SF, Molenberghs G, Sylvester RJ. Identifying the types of missingness in quality of life data from clinical trials. *Stat Med* 1998, **17**, 739–756.
17. Molenberghs G, Kenward MG, Lesaffre E. The analysis of longitudinal ordinal data with non-random dropout. *Biometrika* 1997, **84**, 33–44.
18. Diggle P, Kenward M. Informative drop-out in longitudinal analysis. *Appl Stat* 1994, **43**, 49–93.

19. Curran D. Analysis of incomplete longitudinal quality of life data, Ph.D. thesis. Belgium, Limburgs Universitaire Centrum, 2000, 92–122.
20. Troxel AB, Fairclough DL, Curran D, Hahn EA. Statistical analysis of quality of life data in cancer clinical trials. *Stat Med* 1998, **17**, 653–666.
21. Troxel AB. A comparative analysis of quality of life data from a Southwest Oncology Group randomized trial of advanced colorectal cancer. *Stat Med* 1998, **17**, 767–779.
22. Fairclough DL, Peterson HF, Cella D, Bonomi P. Comparison of model based methods dependent on the missing data mechanism in two clinical trials of cancer therapy. *Stat Med* 1998, **17**, 781–796.
23. Zee BC. Growth curve model analysis for quality of life data. *Stat Med* 1998, **17**, 757–766.
24. Curran D. Analysis of incomplete longitudinal quality of life data, Ph.D. thesis. Belgium, Limburgs Universitaire Centrum, 2000, 158–181