Pergamon

European Journal of Cancer, Vol. 34, No. 9, pp. 1317–1333, 1998

Printed in Great Britain
0959-8049/98 $19.00+0.00

PII: S0959-8049(98)00074-4

Special Paper

Methodological and Statistical Issues of Quality of Life (QoL) and Economic Evaluation in Cancer Clinical Trials: Report of a Workshop

N. Neymark,[1] W. Kiebert,[2] K. Torfs,[3] L. Davies,[4] P. Fayers,[5] B. Hillner,[6] R. Gelber,[7] G. Guyatt,[8] P. Kind,[9] D. Machin,[10] E. Nord,[11] D. Osoba,[12] D. Revicki,[13] K. Schulman[14] and K. Simpson[15]

[1]EORTC Data Centre, Av. E. Mounier 83/Bte 11, Brussels, Belgium; [2]Medtap International, Gilbert Street, London, U.K.; [3]Janssen Research Foundation, Turnhoutsebaan, Beerse, Belgium; [4]Centre for Health Economics, University of York, York; [5]Medical Research Council, Cancer Trials Office, Cambridge, U.K.; [6]Medical College of Virginia, Division of General Internal Medicine, Richmond, Virginia; [7]Dana-Farber Cancer Institute, Division of Biostatistics, Boston, Massachusetts, U.S.A.; [8]McMaster University, Department of Clinical Epidemiology and Biostatistics, Ontario; [9]Royal Victoria Hospital, Division of Clinical Epidemiology, Montreal, Quebec, Canada; [10]Medical Research Council, Cancer Trials Office, Cambridge, U.K.; [11]National Institute of Public Health, Oslo, Norway; [12]BC Cancer Agency, Vancouver, Canada; [13]Medtap International, Bethesda, Maryland; [14]Georgetown University Medical Center, Washington, District of Columbia; and [15]Chiron Diagnostics, Emeryville, California, U.S.A.

In recent years, quality of life (QoL) and economic evaluations have become increasingly important as additional outcome measures in cancer clinical trials. However, both fields of research are relatively new and in need of finding solutions to a substantial number of specific methodological problems. This paper reports on the proceedings of a symposium aimed at summarising and discussing some of the most contentious methodological and statistical issues in QoL and economic evaluations. In addition, possible solutions are indicated and the most pertinent areas of research are identified. Issues specific to QoL evaluations that are addressed include clinically meaningful changes in QoL scores; how to analyse QoL data and to handle missing and censored data and integration of length of life and QoL outcomes. Issues specific to economic evaluations are the advantages and disadvantages of various outcome measures; statistical methods to analyse economic data and choice of decision criteria and analytical perspective. How to perform QoL and economic evaluations in large and simple trials and whether the gap between QoL and utility measures can be bridged are also discussed.

Key words: cancer, quality of life, economic evaluation, costs
Eur J Cancer, Vol. 34, No. 9, pp. 1317–1333, 1998

INTRODUCTION

THE EORTC organised a symposium on statistical and general methodological issues in quality of life (QoL) and economic evaluation in connection with clinical trials in cancer. The symposium took place in Brussels in December 1995 and was supported by the 'Europe against cancer programme' of the European Commission (see Acknowledgments).

Correspondence to N. Neymark.
Received 25 Mar. 1997; revised 2 Feb. 1998; accepted 20 Feb. 1998.

The major objectives of the symposium were:

(1) To summarise and discuss contentious methodological issues in QoL and economic evaluations of anticancer therapies.
(2) To indicate possible solutions to the methodological issues raised and to identify areas where further research is necessary. The aim was *not* to pretend to come up with 'definite solutions', but to summarise divergent opinions and their practical implications. The 'second best' solutions proposed should reflect

the divergent opinions and be susceptible to future revision after critical review.

(3) To publish the proceedings of the symposium so that it can serve as a guide for non-experts with a concise statement of the problems and the provisional solutions suggested.

The meeting was attended by 40 experts from Europe and North America (Appendix). These had been invited on the basis of their publications and significant contributions to the development of the fields and their demonstrated insight in the methodological problems to be discussed. Another criterion was that the participants should primarily be Europeans, as the workshop was placed in a European setting and with financial support of the European Commission. Approximately one third of the participants could be categorised as mainly economists, statisticians or QoL analysts, but there was considerable overlap between these categories. A number of subjects for the sessions of the meeting had been specified in advance (see Table 1) and for each session one of the participants had been asked to write a short background paper, which should serve as a starting point for the discussion. The background papers reflected the author's personal interpretation of the particular subject, the main points of contention and, where possible, indicated proposals for a pragmatic approach to solve the problems.

In this paper, we report on the proceedings of the symposium. The individual sessions are reported separately, with a brief introduction to the main subject, a summary of the background paper (or other material) provided and a synthesis of the main points of the discussion.

It is impossible to reflect accurately all the opinions expressed in the discussions in a paper like this. Parallel and common sessions combined comprised 24 h of intense debate and all sessions were tape-recorded and subsequently transcribed. The parts of the present paper giving an account of the discussions can evidently only give a selection and interpretation of the actual debate.

We had hoped to conclude this paper with a section containing clear guidelines and proposals for solving the problems discussed, based on a consensus resulting from the discussions. The failure to reach a consensus position on most of the issues is a reflection of the complexity of the methodological and statistical problems in these research areas and maybe also of vested interests of researchers involved in this type of research for many years. Hopefully, the reader will benefit from this guide to the essentials of the methodological problems discussed, despite not being presented with consensual solution proposals.

A number of pertinent references have been added. These references will point the interested reader to studies of particular interest, such as exemplary investigations defining and discussing main points of the discussion or primarily methodological papers.

Table 1. *Subjects of the sessions*

Issues for the sessions on quality of life (QoL) evaluation
Clinically meaningful changes in QoL scores
Methods of analysis
Methods of sample size calculation if QoL is a main endpoint
Handling missing and censored data
Reporting results
Integration of length of life and QoL outcome measures
Issues for the sessions on economic evaluation
Clinical trials as a vehicle for data collection in economic evaluations
Relative strengths and weaknesses of various outcome measures
Statistical issues in analysing economic data from clinical trials
The choice of decision criteria and analytical perspective
Can 'rules' for good modelling be established?
Issues for common sessions
(How) can the gap between psychometric QoL outcome measures and utilities be bridged?
How can QoL and economic evaluations be carried out in 'large and simple trials'?

CLINICALLY MEANINGFUL CHANGES IN QoL SCORES

Since sound instruments for measuring QoL have been developed (some examples would be the Sickness Impact Profile [1], the Nottingham Health Profile [2], the SF-36 [3], the McMaster Health Index Questionnaire [4] and the EORTC QLQ-C30 [5]) it is now increasingly being measured for different reasons, with various applications and in a large variety of patients. A problem that is becoming paramount in comparative and predictive studies is how to interpret the results of QoL assessments and how to classify magnitudes of changes. An important phenomenon in this respect is that 'statistically significant' changes are not by definition the same as 'clinically significant' or even 'clinically meaningful' changes.

Background material (presented by Gordon Guyatt)

Whereas psychologists and psychometrists have pioneered the developments of instruments to measure human attitudes, values and experience and health economists have provided an alternative framework for thinking about the impact of health interventions on people's lives, there is not yet a well-established methodology to determine the interpretability of the scores of an instrument. Interpretability means the extent to which we understand what differences in scores constitute trivial, small, medium, or large treatment effects. This aspect is closely related to the property of responsiveness or sensitivity of an instrument.

A number of approaches for establishing an instrument's interpretability have been described in the literature. The longest-standing approach is that of examining effect size. Effect size is calculated by dividing the observed differences in mean scores of treatment and control groups by the standard deviation of the treatment group. The fundamental problem with this approach is that the between-person standard deviation depends on the heterogeneity of the population. If a trial enrols a very heterogeneous population, an important effect may be small in terms of between-person standard deviation and subsequently judged trivial. Alternatively in a very homogenous population, the same effect size may be large in terms of between-person standard deviation and, thus judged extremely important. In these cases, the interpretation differs greatly, whereas the true impact of the change is the same. The significance and interpretation of a given change should be the same, irrespective of the composition of the population.

Other approaches to enhance the interpretability of an instrument are based on the principle of comparing its results with a credible and independent standard that is itself interpretable. One way to achieve this is to use patients' judgments to investigate within-patient differences. This can be done by

asking patients to provide a global rating of change in their status on certain QoL dimensions. Changes can be expressed on a 7-point scale in two directions (improvement and deterioration) and a middle score of 0 representing no change. Thus, the total global rating scale for changes consists of a 15-point scale. Changes can be classified as follows: changes of -3 to -1 or $+3$ to $+1$ are small changes representing minimally important differences; changes of -5 and -4 or $+5$ and $+4$ are moderate changes; and -7 and -6 or $+6$ and $+7$ are large changes. Combining and averaging associated domain-specific scores provides a mean score of change for patients with a particular disease. Very consistent results have been obtained using this technique across highly varying diseases. The consistent pattern is that a difference of 0.5 per disease-specific question is a minimal difference; a difference in the range of 0.75–1.25 can be considered a moderate difference; and a difference of more than 1.5 can be considered as large.

Instead of looking at within-patient differences, one can also focus on between-patient differences. In this approach, patients with a particular condition are asked to rate themselves relative to another person with whom they have been speaking. Subsequently, the patients are asked to fill out the QoL questionnaire and a definition of difference is derived from the comparison of the rating and the questionnaire. A general feature looking at interpersonal differences is that, after speaking to other patients, people tend to rate their own situation a little better than that of the person they are comparing themselves with. This results in a shift upwards of all scores. However, this can be compensated for by taking the average scores of all who state that they are equal and comparing this with the average scores of all who state that they are a bit better.

Instead of using patients' ratings one could also use clinicians' judgments of changes in scores of measures that are well known to them. For instance, one could administer questionnaires to patients before and after an intervention of known effectiveness with which clinicians are familiar, so that they can see the change in score associated with response to treatment. Other possibilities include the comparison with elements of life to which we all can relate, like physical functioning. For instance, one could relate changes of physical well-being to an improvement from moving about in a wheelchair to walking with physical limitations.

To summarise, the evaluation of clinically meaningful changes requires three conditions: firstly, the need of an independent standard which is itself interpretable; secondly, the change must be meaningful to the audience who use the information, i.e. the clinicians; thirdly, the instrument must bear at least moderate correlation with the independent standard. At present there is no well-established methodology that solves this problem of clinically meaningful changes with a single strategy. Therefore, it is necessary to use a variety of approaches to increase the meaning of QoL measures.

Discussion

It was stressed that the above-mentioned approaches have (so far) not been applied in cancer clinical trials. An approach that has been developed by the National Cancer Institute of Canada is a subjective significance module of questions to investigate patients' perception of change on four of the domains of the EORTC QLQ-C30 questionnaire. Each question asks patients to compare their current situation with the previous occasion at which they filled out the questionnaire. Answering categories are a semantic differential 7-point scale.

There may be a problem related to this method because of the inevitable rephrasing when asking if a change has occurred on a certain domain. The subjective significance term differs from the phrasing of the individual items in the QLQ-C30 questionnaire. As a result, one is not sure whether the patient attributes the same meaning to such a term and, hence, there is uncertainty over the congruency of the two types of scores reflected by weak correlations between the two measures. The NCIC found that the physical functioning item of the subjective significance scale and the physical functioning scale of the QLQ-C30 had a correlation of only 0.23, whereas the social item and scale had a correlation of 0.42. However, it was emphasised that a low correlation may be due to other factors as well (e.g. insufficient variability in a population or in the magnitude of change).

The problem related to the use of a descriptive score and a subjective change score is, therefore, which score to use as the gold standard? One argument for focusing on the descriptive QoL scores is the difficulty in adequately memorising the situation of the previous assessment, if the interval between the two assessments becomes long. Moreover, there is also the process of adaptation to the situation. The process of adaptation could even result in a reconsideration of the perception of a previous situation. Alternatively, if we agree that the patient is the best and final arbiter of his/her health outcome, then it can be argued that these subjective evaluations are in fact the optimal information to use. Those in favour of this point of view suggested dropping the term 'clinical' in the discussion of meaningful changes, because it implies that medical parameters are more important than the view of the patient. Finally, a kind of midway position was agreed upon, that as long as we do not yet have a clear understanding of the patterns in subjective significance score and the dynamics of QoL as measured by the standard psychometric approach, it seems best to incorporate all available data in the process of medical decision making, not only patient-derived data, but also clinicians' judgements.

Future research to improve methods of determining clinically important changes could be to develop associated questionnaires of subjective significance and relate these to other more objective external indicators of change, such as response to treatment, or formal diagnostic systems, like those used in psychiatry and clinical psychology. A possibly useful method to improve response correctness could be to show patients the scores of their previous assessment. So far, only one group of researchers have investigated this and the results seemed promising. However, this needs to be replicated by others. A final approach worth developing is a mapping system of reference scores of all kinds of health states throughout a large variety of diseases.

STATISTICAL APPROACHES TO ANALYSING QoL DATA

During the past two decades of QoL research, the main emphasis has been on the development of sound measures to evaluate QoL. Issues of statistical analysis of QoL data have received little attention in the literature so far. Now that good tools to measure QoL are available, it has become clear that there is a pressing need to find proper methods for analysing QoL data and to gain experience using them.

Background material (presented by Peter Fayers)

A randomised clinical trial is the only form of scientific investigation that can ensure an unbiased comparison of the efficacy of different treatments and QoL assessment is an essential component of this treatment evaluation. However, there are many problems regarding methods of analysis and interpretation of the results. Analysis of QoL data from cancer clinical trials can be classified into two broad categories: confirmatory data analysis and descriptive or exploratory data analysis. Confirmatory data analysis is used when a number of a priori hypotheses are formulated and tested. The hypothesis testing can be largely based upon standard statistical significance testing, although there may be practical problems arising from the multidimensional nature of QoL assessments, the 'longitudinal' nature of the repeated measurements over time and the occurrence of missing data for individual patients. Exploratory and descriptive data analyses, as their name suggests, are used to explore, clarify, describe and interpret the QoL data. Frequently, these analyses will reveal unexpected patterns in the data, for example suggesting differences in QoL with respect to treatment or other factors. However, exploratory analyses often consist of a large number of individual comparisons and significance tests and some apparently strong effects may in fact arise out of chance fluctuations in the data (there is considerable variability and 'noise' when assessing QoL). Thus, exploratory analyses may result in the generation of new hypotheses which should then be in subsequent confirmatory studies.

Because exploratory and descriptive analyses are less concerned with significance testing, graphical methods may be especially suitable. These have a number of advantages over purely numerical techniques. In particular, judicious use of graphics can succinctly summarise complex data which would otherwise require extensive tabulations and can clarify and display the complex interrelationships of QoL data. At the same time, graphics can be used to emphasise the high degree of variability in QoL data. This contrasts with numerical methods, which may often lead to results being presented in a format which leads readers to assume there is greater precision than the measurements warrant. Graphical techniques can highlight changes in QoL which are large and clinically significant, whilst making it clearer to readers which changes are unimportant (even though some clinically unimportant changes may be statistically significantly different from zero).

QoL measurements are frequently collected at repeated times before, during and after treatment. Two of the most basic methods of graphical display of QoL over time are plots of mean scores over time and plots over time of the percentage of patients with values exceeding a certain level. For example, in considering an assessment of pain (say, a 4-point scale with 1 = 'not at all', 2 = 'a little', 3 = 'quite a bit', 4 = 'very much'), one might either plot the mean pain score over time, or the percentage of patients reporting serious pain (e.g. pain categories 3 and 4). The percentage plots could be extended by superimposing plots corresponding to the percentage of patients in each of the four categories.

In summary, most methods of analysis and interpretation are susceptible to problems of bias. The use of numerical methods and formal hypothesis testing may lull readers into a false sense of security, since statistical significance might be falsely assumed to imply that the hypotheses have been confirmed or rejected. Graphical methods, by being less formal, may be more appropriate and may more readily lay emphasis upon the variability and uncertainty attached to QoL measurements.

Discussion

There was general agreement that the nature of QoL data does not require the invention of new statistical methods to analyse the data. Issues like time dependency and repeated measures can, in principle, be handled using conventional statistical methods. However, there are some other obstacles that often necessitate a pragmatic approach to the analysis of QoL data.

The first obstacle is the lack of agreement about the best approach. The choice of the best approach depends not only on the research question, but also on the use of the results. Are the results to be used for decision making at an individual patient level or for decisions regarding the optimal treatment for groups of patients? For decision making at an individual level, descriptive information presented in a graphical way seems to be the most appropriate. If QoL is the primary endpoint in a randomised clinical trial comparing two different treatment strategies, the appropriate statistical methods would be those that allow confirmatory data analysis. Different decisions require different types of information.

There was some general concern among the participants about the inappropriate use of statistical methods to analyse QoL data for the sake of comprehensibility and interpretability of the results. Clinicians prefer to read an overall conclusion of QoL results, preferably accompanied by a *P* value. Very often, the available data are such that the underlying assumptions for applying those types of statistical analyses are not met.

Two opposing views became apparent during the discussion. One faction was of the opinion that at present only simple and conservative methods of descriptive data analysis should be applied, with an emphasis on graphical methods. They stated provocatively that testing of a priori formulated hypotheses is questionable as the principal aim of QoL data analysis. A description and comprehensive interpretation of patterns of change would be more appropriate. The opponents argued that this would imply the loss of much information. They recommended to start data analysis at the point where the other group stops analysing data and favoured further analysis despite the problems caused by missing data. According to this viewpoint, it would be very unsatisfactory to perform a clinical trial in which QoL is an endpoint and ending up with only descriptive results in terms of percentages without even an attempt to answer the initial question about the differences in length of life and QoL with the treatments under comparison.

A second obstacle that necessitates the use of pragmatic solutions to analyse QoL data is the current lack of large databases. As long as we do not know what 'normal' values are for QoL in particular patient populations, any decision on what a meaningful change is remains arbitrary. As we accumulate more information on the score levels for different groups of patients, we may become able to determine what constitutes a meaningful change. It was generally agreed that the use of sensitivity analysis and confidence intervals should be strongly recommended. Also, that it is important to provide a full description of the characteristics and problems of the available data, as well as a description of the likely direction of the bias of the results.

A third obstacle is the problem with missing data. Although the best way to deal with missing data is not to have any, missing QoL data are very common in practice. The problems arise at the stage of data collection and must necessarily be addressed at this stage. However, missing data are, and to a certain extent will always remain, an inherent characteristic of a QoL dataset. The type and amount of missing data determine to a large extent the type and robustness of the conclusions that can be drawn. Handling longitudinal QoL assessments with large amounts of missing data remains problematic.

SAMPLE SIZE CALCULATION AND HOW TO HANDLE MISSING AND CENSORED QoL DATA

With respect to the determinants of sample size, norms and precedence dictate values for power and significance level, while the crucially important anticipated effect size must be determined for each trial, based on experience, published data or pilot studies. With QoL as a main outcome of a trial, there is neither experience of nor agreement about what would constitute a meaningful benefit for the patient. Missing data can imply a significant loss of power of a study. The cumulative effects of many missing values can drastically reduce the number of patients available for analysis. Another problem is the possibility of serious bias, if the data are not missing at random.

Background material (presented by David Machin and Peter Fayers)

In a phase III clinical trial, it is necessary that there is an adequate sample size to provide sufficient power to test the significance of differences in treatment effects. Historically, the main endpoint in cancer clinical trials has been clinical (e.g. survival, disease-free survival or response rate) and the sample size calculation has been based on either a time to event analysis or a comparison of two binomials. However, in recent years, several clinical trials have been initiated with a QoL outcome as the main endpoint. A common problem with QoL as a main endpoint in cancer clinical trials is the specification of a worthwhile and clinically meaningful treatment effect size that is plausible in practice. There is some evidence that a worthwhile effect is often larger than the plausible effect, implying that we are over-optimistic in stating what the treatment effect is that we should be looking for. Another major problem has been that many clinical trials have been under-sized and under-powered, leading to a decreased chance of detecting differences between treatments. Earlier cancer clinical trials were designed to show a difference between treatments, while today's trials tend to be designed to test equivalence. This use of the reverse of the null hypothesis has important consequences for the sample size needed.

Missing data are usually a particularly severe problem in the context of large multicentre randomised trials, especially where the QoL assessments span a substantial period of the patients' follow-up and where the survival rates are low. Missing data are an important problem, because they result in a loss of power. Firstly, they reduce the number of patients that are available for simple univariate analysis. Secondly, many forms of repeated measures analyses or multivariate analyses assume that complete data are available and the cumulative effects of missing values can drastically reduce the number of patients with complete data that are available for such analyses.

Missing data can introduce bias if we do not know whether data are missing at random. If data are not missing at random, what is the impact of ignoring the missing values? And how can one best estimate or impute the missing values? There are some methods of multivariate analysis that allow missing values to be ignored. However, this is usually tantamount to assuming that missing data are missing at random.

Two types of missing data may be identified. Firstly, there can be items missing within a form. This is the simpler example of missing data. However, can one assume that those items are missing at random? In many cases, it may appear likely that the patient simply forgot to answer all the questions, in which case one may be tempted to impute the missing value using simple estimation procedures. In other cases, it is perhaps more likely that the patient consciously avoided answering the question(s).

The second type of missing data is that of missing forms. When a whole form is missing the problems are similar but more extensive. Missing forms are, in many trials, more common as the patients come closer to death. So missing forms occur frequently during later stages of clinical trials. In this situation it seems likely that the forms are not missing at random.

So far, there is no consensus on guidelines in the case of missing items or forms. Each case has to be reviewed on an individual basis. In this context, it is of crucial importance to gain information on the possible reasons for missing data. There are three main methods for handling missing data:

(a) Regard each assessment separately and plot all available data at each time point. Thus, one might plot or tabulate the results across all patients at the baseline time point, using all completed baseline forms, and in the same analysis use all available forms for each of the other time points. This leads to different patients being used at each time point, according to whether or not they completed the questionnaire at that time. The main advantages of this method are its simplicity and the fact that it uses all available data. However, there are potentially serious problems, since different patients and different numbers of patients are used at each time point.The comparisons across assessments may thus be biased.

(b) The main alternative method is to restrict the data set to those patients who have returned complete data at all measurement time points. For example, if one wishes to analyse patterns of QoL during and immediately post-treatment, one might define the period of interest as the first 6 months. Then the data set would comprise those patients who have returned forms on all occasions during the first 6 months. One might additionally adapt this approach to allow for censoring due to death. This method, too, has an advantage of simplicity. However, the sample size available for analysis may become greatly reduced, since different patients may have missing data at different times. More seriously, there remains a problem of bias because the patients that survive and have good compliance may well have better physical and psychological status than other patients. Moreover, if survival rates are differentially affected by the treatments, there

will be different numbers of patients included in the QoL analysis of the two treatment groups. This could represent unacceptable bias.

(c) The third method is to 'impute' values for the missing data. In principle, a model is developed corresponding to the pattern of missing data and this is used to estimate the most likely values whenever data are missing. The mechanics of imputation are relatively straightforward—either by direct estimation using averages or regression techniques, or by an iterative method such as the expectation maximisation algorithm. However, in all cases of imputation the most crucial aspect is the specification of an appropriate model for the pattern of informative or non-random missing data.

Discussion

During the discussion, the importance of the distribution of scores and criteria for interpreting changes in QoL scores once again became apparent. It was stressed that the sample size calculation should be based on the type of data (categorical, ordinal or cardinal) and the method of analysis that will be applied (parametric or non-parametric). To perform sample size calculations we require estimates in the control arm. As QoL data are accumulating rapidly, the problems in calculating sample sizes should become less.

Sample size calculations are usually made based on the primary endpoint. Fortunately there have been only very few clinical trials in which QoL was the primary endpoint. There was some general agreement that for QoL as a secondary endpoint the sample size of the primary endpoint would and should be large enough to provide adequate power to detect substantial differences between the two treatment arms with respect to QoL. The general advice is to base sample size calculations on only a primary domain of QoL instead of the whole spectrum.

With respect to missing QoL data, there was a general consensus that surrogate data are preferable to no data at all. Various suggestions were made on how to impute data, such as proxy ratings, latest scores taken forward, impute extreme values and perform a sensitivity analysis. It was pointed out that a number of new methods relying on more sophisticated modelling techniques are currently under development or being tested. All these techniques depend greatly on the validity of specific assumptions and their value still has to be documented.

INTEGRATION OF LENGTH OF LIFE AND QoL OUTCOME MEASURES

Length of life outcomes have been classical endpoints in cancer clinical trials. Now that QoL is increasingly included as an endpoint, there is a need for special methods to combine length of life and QoL data leading to a valid measure of quality adjusted survival. Well-established methods are not yet available, but in recent years a special form of quality adjusted survival has become quite popular for use as a decision aid in resource allocation problems. This is the quality adjusted life years (QALY) approach, but there are some major unsolved theoretical and methodological problems related to this technique. An alternative, the quality adjusted time without symptoms and toxicity (Q-TWiST) model, is a method especially developed for integrating QoL adjustments in survival analysis.

Background material (presented by Richard Gelber)

The evaluation of treatments in cancer clinical trials utilises a variety of endpoints, including length of life outcome measures (overall survival, disease-free survival), but also data on toxicity and QoL. For chronic illnesses with no cure, new treatments should be evaluated not only for their effect on survival, but also for possible palliative advantages. A given therapy will not often be equally advantageous with respect to all outcomes considered important and it is in general important to gain insight in the trade-offs between the various outcomes. The most important case is the frequent trade-off between a gain in survival and increased toxicity (at least in the short run) and its presumed detrimental effect on QoL.

Quality adjusted survival analysis may be used as a general term for methods to take such trade-offs into account. As in an ordinary survival analysis, the focus of quality adjusted survival analysis is on time, but survival time is adjusted to take account of negative impacts (e.g. of toxicity) on QoL. The adjustment is accomplished by a kind of weighting system, with weights ranging from zero (a state of health assessed as equally as bad as death) to 1 (perfect health, in some sense). Conventional survival analysis implicitly assumes the weight 1 for the whole survival period, while quality adjusted survival analysis takes into account that QoL may vary over time, depending on the effects of treatment and other factors and assumes that QoL may be measured meaningfully on an interval scale from 0 to 1. A composite measure of quality and quantity of life (i.e. quality adjusted survival) is obtained by summing the weighted periods of survival time. This summary measure can be used for treatment comparisons and as the outcome measure in cost-utility analyses. There are two main variants of quality adjusted survival analysis, the QALY method and the Q-TWiST method.

The QALY method. In a QALY analysis, specified time intervals are set prospectively based on the time points at which utility assessments are scheduled. The total QALYs following a treatment can be computed per individual patient as the sum of the products of each assessed utility score multiplied by the duration of the time period for which this utility measure is assumed to be valid. Subsequently, the QALYs for all patients in a given treatment arm are averaged to provide a QALY estimate for that therapy. A good example of the use of QALYs in a cost-utility analysis is provided by [6]. The US panel on cost-effectiveness in health and medical care [7] recommends using QALYs for the reference case analysis and, in addition, contains an extensive discussion of the issues in selecting such outcome measures (e.g. QALYs versus healthy years equivalents) and in measuring these outcomes.

QALY analyses have not been performed frequently in cancer clinical trials because of several limitations. First, the serial utility assessments require lengthy individual interviews that are often not feasible. A second limitation is the problem which arises from missing data in longitudinal assessments. A third limitation for calculating QALYs in clinical trials relates to the handling of censored observations. If a Kaplan–Meier estimator is applied directly on the accumulated QALY score for each patient up to his or her follow-up time, this estimate will be biased, since the distribution of QALY scores and the censoring are not statistically independent.

The Q-TWiST method. In a Q-TWiST analysis the course of the patients' treatment and follow-up is partitioned into

health states that are clinically relevant for the therapies under consideration and a utility score is assigned to each clinical health state. Instead of forming quality adjusted survival estimates from each individual's history, the average amounts of time patients spend in QoL oriented clinical health states are estimated and these are combined as a weighted average with the utility scores as weights to form the Q-TWiST estimates. This procedure can be used with censored data without the difficulties associated with informative censoring described above.

The Q-TWiST method utilises a three step procedure

The first step is to define progressive QoL oriented clinical health states that highlight differences between the treatments being compared. For example, in the case of adjuvant chemotherapy for resectable breast cancer, the time with toxicity (TOX) is represented by the period in which the patient is exposed to subjective side-effects of therapy. The time without symptoms of disease or toxicity of treatment (TWiST) is a state of relatively good QoL. The time spent living with overt metastatic disease or time in relapse (REL) represents the time after the diagnosis of systemic spread of the disease until death.

The second step is for each treatment separately to partition the overall survival time into the duration of the clinical health states by using data for the time points of transition between the states defined in the first step. For example, areas between the Kaplan–Meier estimates for overall survival, disease-free survival and time to end of toxicity represent the average amounts of time spent in the respective clinical health states: TOX, TWiST = disease-free survival-time to end of toxicity and REL = overall survival—disease free-survival. These are calculated up to a specific point in time determined by the follow-up limits of the study cohort. The resulting estimates are called restricted means because they represent the mean survival time within the follow-up interval. The survival curves for the outcomes can be plotted on the same graph to illustrate the partitioning according to treatment group. This is known as a partitioned survival analysis.

The third step is to compare the treatment regimens in terms of quality adjusted survival. This composite measure is obtained by summing the average clinical health state durations calculated in the second step multiplied by utility coefficients. TWiST is considered to be the best possible clinical health state for the particular disease being studied. In most analyses it is assigned the value of 1 for all treatment arms. When patient-derived utility values are not available, the treatment comparison results are best presented as a threshold utility analysis. This is a sensitivity analysis comparing the treatments over all possible combinations of the utility scores for the clinical states (TOX and REL in the example), whereby the threshold utility values changing the preference from one treatment to the other (threshold) may be determined.

The Q-TWiST method has several advantages as a quality adjusted survival methodology. It can be applied retrospectively even without patient-derived QoL measures. Its graphical presentations are understandable to clinicians who are familiar with Kaplan–Meier survival curves. The method works with censored data. The threshold utility plots allow patients to identify which treatment they may prefer, depending on their individual constellation of relative utility values for TOX and REL.

Discussion

The discussion focused mainly on the perceived commonalities and differences of the QALY and Q-TWiST approaches. For some, the two approaches are conceptually quite similar, whereas others consider them to be very distinct approaches. However, most participants agreed on the following three points.

(1) The underlying theoretical assumptions of the approaches are similar.
(2) Both approaches have the same objective, i.e. to integrate length of life and QoL outcomes by adjusting the time spent in a certain health outcome for the quality of that health state.
(3) Both methods include the same initial basic elements of a stepwise procedure: the relevant health states have to be defined; the time spent in each health state has to be calculated or estimated, and a value has to be assigned to the various health outcomes.

However, the calculation of outcome presents a major distinction between the two methods. The end product of the QALY method is a single number representing the estimation of life years gained. The outcome of this method can be interpreted and compared over studies. In the Q-TWiST method, the final outcome is a sensitivity and threshold analysis which is context dependent. The fact that one method is context dependent and the other is not, is due to their respective approach of assigning weights to possible outcomes. In the QALY approach, the anchors are fixed: the lower boundary represents death and is assigned the value of zero and the upper boundary is perfect health, which is assigned the value of 1. In the Q-TWiST, the maximum value is the best possible health outcome that can be obtained in a certain context. The other health states are valued relative to this. So, the weights used are not necessarily absolute weights, but relative to the best possible health state within that specific situation.

Another important distinction between the two methods is their areas of application. The QALY method is mainly used in cost-effectiveness analysis and decision making at a policy level requiring the use of a single outcome measure. The Q-TWiST method is used in the context of clinical trials where different treatments are compared for their efficacy. In this case, it is not necessary to obtain a single figure to make decisions about the optimal strategy for groups of patients.

During the discussion it also became apparent that some clarification was needed with respect to the Q-TWiST method. One issue to be clarified was the perception that the Q-TWiST method can only be applied in an adjuvant setting. Although the method was initially developed for adjuvant treatment of breast cancer, it is currently also proposed in non-adjuvant settings and for other diseases, for instance in metastatic lung cancer. In such a study, the period of TWiST is likely to be very small or even non-existent. It could be replaced by a best possible health state which is specific for that study.

Another issue that needed clarification is the maximum number of health states. If there is evidence that there can be various relevant health states with different associated values, then these health states have to be considered in the model. For instance, long-term side-effects could be a relevant issue to consider separately in the model. The inclusion of more

health states would certainly complicate the analysis, but it is feasible.

One difficulty associated with the Q-TWiST method is its retrospective application. The definition of health states and the transitions between health states once the survival benefits are known may bias the results. A prospective definition of health states and transitions prior to the start of the study is to be preferred over a retrospective evaluation. However, for some purposes a retrospective analysis can be warranted. In this case, the definition of the health states and transitions have to be carried out in a conservative way with respect to the medical outcomes.

Finally, two problems related to the method were mentioned. The first is the definition of the health states. Health states are not static, but tend to vary not only between individuals but also within individuals over time. In particular, the toxicity experienced by patients during treatment may vary substantially across patients and over time. So, the definition of the health states may not adequately cover all relevant aspects at all times and for all individuals. Although relevant, this problem is not unique to the Q-TWiST method. It applies to any method that uses descriptions of health outcomes, which always imply a simplification of the complexity of the real world.

The second problem concerns how to obtain the utility scores for the various health states. In the early studies in which the Q-TWiST method was developed, arbitrary utility values were used, just to illustrate the workings of the threshold analysis technique. The use of formal methods (such as the standard gamble or time trade off) to elicit patient utilities is always to be preferred over the assignment of arbitrary values, but is rarely feasible, as these techniques are not easy to understand and are very labour intensive. For these reasons they are difficult to use in a clinical trial. However, it was concluded that this problem is not unique to the Q-TWiST method either.

A BRIEF INTRODUCTION TO ECONOMIC EVALUATION

The term economic evaluation refers to a general methodological approach to assess the relative value of goods, services or activities. For goods traded in normal markets, this valuation of the relative value of the various possibilities is carried out by the buyers when they consider the acquisition of one or the other. The use of formal economic evaluation as an aid to decision making becomes relevant for goods and services for which the normal market mechanisms, for one reason or another, cannot be assumed to function well. Health care is one important area of the economy, where, for a variety of reasons, the usual market forces are not expected to function well and public regulation is the rule rather than the exception.

The following is a brief, general definition of economic evaluation as a method of analysis: *a comparative analysis of alternative lines of action including their effects as well as their costs*. The basic tasks in any economic evaluation are, therefore, to identify, measure, value and compare all the costs and effects of the alternatives under consideration. The purpose is to carry out a structured collection of information regarding each of the alternatives and to ensure that the information about all the alternatives is collected, treated and assessed according to uniform and consistent criteria. Applied to medical care, the methods of economic evaluation may be used to compare alternative treatments of specific diseases or, more broadly, to compare the contribution of various (programmes of) interventions with an overall objective, such as improving (quality adjusted) survival. The basic reference in this field continues to be [8], while the report by the US panel on cost-effectiveness in health and medicine mentioned above [7] may become the new standard reference.

Various types of economic evaluations are encountered in health care. The methods diverge mainly in the way they assess and value the effects of the interventions on the patients' health status. The most commonly used type of analysis is cost-effectiveness analysis, which requires that the effects of the interventions be expressed in a single clinical dimension by means of a 'natural unit' of measurement, such as life years gained or reduction in probability of relapse for cancer patients. If the treatments' effects can be considered to be essentially identical, a cost-minimisation analysis may be used. The assumption of identical effects is not often defencible, although it may sometimes be justifiable to focus the analysis on a single dimension of outcome (e.g. survival in patients suffering from acute life-threatening diseases). Cost-utility analysis can be considered as an extension of cost-effectiveness analysis, aiming to overcome the limitations caused by the requirement that outcomes should be assessed in one single dimension. By assessing the net effect of treatments on patients' utility or health-related QoL, treatment outcomes are measured in terms of patients' or society's perception of treatment benefit. Finally, in cost–benefit analysis treatment outcomes are valued in pecuniary terms, whereby benefits and costs are rendered comparable and the net benefit of each of the comparators may be calculated. However, this method is used infrequently because of the difficulties posed by the pecuniary valuation of outcomes.

The first economic evaluations of health care interventions appeared approximately 25 years ago, and the number of studies published each year has been growing steadily. Such analyses are no longer merely academic exercises, but are increasingly being used to justify, if not determine, decisions in everyday medical practice. For instance, the decision may concern the inclusion or not of a new drug on a hospital or primary practice formulary listing or its reimbursement status. The increasing importance of such analyses for practical decisions require accuracy and attention to detail and this will probably give even further impetus to the important discussions of methodological issues, which have always characterised this field of academic inquiry [9–17].

RANDOMISED CLINICAL TRIALS AS A VEHICLE FOR DATA COLLECTION

Economic evaluations of treatment alternatives require valid and reliable data on both effects and costs of each of the comparators. Generally, there is a lack of pertinent data at the time when initial reimbursement decisions must be made, and, habitually, economic evaluations have been performed by bringing together data from a number of different sources. Since its first launch in the mid-1980s, the idea of collecting cost data in connection with randomised clinical trials, usually considered the gold standard for assessing the efficacy of (new) treatments, has attracted increasing interest. However, attention should be paid to a number of potential problems in basing economic evaluations on data collected this way.

Background material (presented by Linda Davies)

A main criterion for determining whether to include economic components in a clinical trial is the relevance of the trial in economic terms. Besides quantitatively important anticipated differences in the costs and/or effects of the alternatives, the pertinent considerations include whether it is possible to estimate the benefits associated with the clinical outcomes in a relevant manner for the assessment of the relative value for money of the alternatives. In addition, the decision about integrating the collection of the necessary data in the trial should be governed by the likely variability of the items between patients and about the potential and expected costs of collecting accurate data outside the clinical trial.

The design of the trial should be appropriate for economic evaluation. The first issue is the external validity of the clinical and economic results, which *inter alia* depends on the inclusion and exclusion criteria and the specification of patient management in the protocol. It may also be difficult to generalise the procedures followed, the resource utilisation and clinical results to routine clinical practice, because trials tend to be conducted by highly committed and specialised investigators in specialist centres. Another design issue is the length of follow-up, since the planned follow-up for clinical trials is typically shorter than the time horizon for an economic evaluation. This may preclude measurement of the resource use and consequences of certain clinical events of importance for the overall costs and effectiveness of an intervention. It also often forces the economic analyst to use surrogate rather than final endpoints. Economic evaluation should always be based on an intention to treat analysis and this requires accurate follow-up even of patients withdrawn from treatment.

Since data collected in clinical trials are stochastic, considerations about appropriate methods of statistical data analysis are called for. This issue is discussed below.

Discussion

A prevailing viewpoint in the discussion was that there are no major theoretical barriers to the inclusion of economic evaluations in clinical trials. One argument in favour is the presumed interaction between resource use and outcome. Another is that it is simply not permissible to combine outcome data from trials with administrative or observational cost data, because the populations are completely different. In general, it is very important that the economists become involved in the planning of the trial from the very beginning and participate in the discussion on the choice of endpoints, length of follow-up, power issues, etc. The case report forms for economic data should always be integrated with the clinical case report forms. The possibility for tracing the patients' use of other medical care services by data from other sources, such as private insurers, were viewed with some skepticism, although such data would be a valuable supplement to the data derived from clinical trials.

Most of the discussion turned on the issue of generalisations from the results of economic evaluations integrated in randomised clinical trials. One viewpoint was rather skeptical about the possibility of generalising the results from clinical trials, viewing the role of economic evaluations alongside clinical trials as very restricted. The opposing view was less reticent and less critical of the relevance of economic evaluations carried out in connection with clinical trials, arguing that economic evaluation within a phase III trial is often the only possibility to assess the economic impact of a new therapy before it is considered for reimbursement [16, 17].

The restrictive viewpoint is that an economic evaluation alongside a randomised clinical trial can at most show whether a new treatment has the potential to be efficient, while it cannot address the question 'will it be efficient in practice?', which is the relevant issue and the one that economic evaluations are intended to inform. There is a clear trade-off between internal and external validity and in a randomised clinical trial one is trying to maximise internal validity in order to establish the efficacy of the new treatment by eliminating, as much as possible, any competing explanations for a difference in treatment outcome. An economic evaluation in a phase III randomised clinical trial is primarily a good means of gaining insight into the principal features and variables that should be included in a subsequent economic assessment based on, for instance, observational data or a non-randomised trial.

Some concern was raised about the possibility of ever having effectiveness or phase IV trials carried out, at least for treatment comparisons showing a clear difference in efficacy, because it would seem unethical to allocate (by randomisation or otherwise) any patients to the less effective treatment. However, this concern was rejected, partly because it ignores the argument for taking both costs and benefits into consideration, simultaneously and with equal weight, partly because the diffusion of new and superior treatments does not take place instantaneously and completely immediately after a significant benefit has been documented once.

There was general concern about the representativeness of centres. This matter is further complicated by the variations between centres in resource use, not just in quantities, but also in types. Can resource uses be averaged when there are such large variations, as observed in several trials and other types of studies, variations due to differences in relative prices as well as to differences in practice patterns, and which may interact with the treatment and outcomes? Another concern was the possible learning effect. As the physicians gain more experience with the new treatment, the suitability of patient types and the handling of side-effects, this could lead to changes in resource consumption.

Concerning the external validity of economic evaluation in randomised clinical trials, it may be claimed that the clinical outcome is the maximum to be expected, simply because more tests are carried out and many things are determined earlier in a clinical trial than outside trials. At the same time, the difference in costs is minimised, because a large part of the cost is protocol driven and there is limited flexibility allowed in the clinical treatment. It is a commonplace to state that protocol driven costs should be excluded from the analysis, but in practice it may be difficult to determine exactly what these costs are. It can be argued that clinical trials usually underestimate the costs in the real world because in everyday clinical practice the treatment will be used in situations where it will not be as effective and this will increase the costs outside the trial. Whether this will affect the comparators to a different degree (and thereby change the estimate of relative costs) cannot be determined a priori, but must be examined empirically on a case-to-case basis.

RELATIVE STRENGTHS AND WEAKNESSES OF THE VARIOUS OUTCOME MEASURES

An essential step in the design of an economic evaluation is the choice of outcome measure. Cost-effectiveness analysis uses outcome measures, such as event rates (e.g. relapses prevented, live years saved), or measures on a health status index (i.e. scores on dimensions such as physical function, pain, social function). Cost-utility analyses use outcome measures which combine duration of life and health-related QoL into a single, numerical summary measure intended to represent the utility for the patients of a set of outcomes. The various outcome measures may be seen as complementary, providing the decision makers with differing types of relevant information.

Background material (presented by Paul Kind)

We are, primarily searching for measures that are sensitive to change, i.e. they must be able to detect real change. We usually rely on repeated measurements of health status or QoL across time and from differences observed imply that a change has occurred. If the measurements are to be used in an economic context, they must be expressed in a quantitative form and have particular arithmetic properties. Measures with ratio scale properties are often sought after, but actually interval scales will do. An interesting case is the frequently used Karnofski index, which is kept in use because previous studies always used it. This is a nominal classification labelled with numbers, so it appears to have arithmetic properties, although it should be considered largely an alpha-numeric, nominal tool. Another case is the Spitzer QoL index, which is just arbitrary weights with an index property, where the weights are basically the product of a factor analysis, which do not necessarily have any basis in patient preferences.

No matter how outcome measures are derived, they are likely to have some underlying methodological components. which will mark the data analysis. A two stage approach to the design and development of the measures is frequently seen. The first is the requirement to describe health status in order to categorise patients and to determine whether they are better or worse, e.g. to determine if a change has taken place. The second step is to find a valuation or weighting system to quantify the change. Psychometricians are likely to favour methods based on ranking, while economists are more likely to prefer choice methods such as standard gamble or time trade-off to obtain utility weights. Whichever method is chosen, there remains the issue of the arithmetic properties of the measure, and the necessity of choosing between the various methods, which will yield different sets of weights.

Discussion

The first issue raised was whether economists should press more for the use of pecuniary valuations, in particular willingness to pay measures instead of, or as a supplement to, the various forms of utility or other outcome measures. One objection to this was that a danger of using pecuniary measures is that people will think they understand them, while they would readily admit that they do not really know what utilities mean.

Another viewpoint was that the chances of ever finding a composite outcome measure that fulfills everybody's objectives are very slight indeed, so it might be better to use separate measures suitable for the varying objectives of the various professions. QoL measures are needed from a medical decision maker's and a patient's point of view, but it may also be useful for economists striving to generate some willingness to pay measures. It is doubtful whether it is useful to try to develop this QoL measure into a quantitative measure to be used for economic evaluations of the cost-utility type and it is probably better to keep a distinction.

One idea could be to ask respondents about their willingness to pay for a change found through a well-validated QoL profile questionnaire. Presumably, the more 'meaningful' a change, the higher willingness to pay. Patient preferences and willingness to pay measures could be obtained outside clinical trials, which should only be used to determine changes in descriptive terms.

One objection to this proposal was that preferences may not be stable over time. For instance, if the patients enter a trial full of hope, over a time period their QoL would change in a negative direction if their optimism fades, even though their health does not deteriorate in an 'objective' sense. Another objection was that it would be a move into a position of the worst possible combination of elements, where aspects of a QoL measure will be singled out and put into scenarios, which patients are then asked to evaluate in some way, while the ensuing information will later be used in quite another context. The procedure would be peculiar to each study, patient preferences and weighting would not be validated, and considerable resource expenditure would be required to obtain the information.

STATISTICAL ISSUES IN ANALYSING COST DATA FROM CLINICAL TRIALS

When the possibility of collecting patient-based data on costs and effects by carrying out economic evaluations alongside clinical trials is being used, the analysis becomes stochastic rather than deterministic, as distributions for both cost and effect can be obtained. The analysis must be conducted with appropriate statistical methods to assess the statistical, as well as the quantitative, significance of differences in costs and cost-effectiveness and new statistical methods may have to be developed to address some issues [18–21].

Background material (presented by Kit Simpson)

Variations in total costs incorporate variability in both quantities and unit cost of the resources used, and this variability is increased in multicentre/multinational trials. The following is a list of some of the most important questions regarding statistical analysis of stochastic cost-effectiveness data currently being addressed by methodological investigators.

(1) What is the magnitude of worthwhile cost differences? This is difficult to specify, but necessary for determination of sample sizes.
(2) Are sampled or non-sampled data on costs more likely to give us valid cost-effectiveness ratios? Are there clear examples that indicate whether a deterministic, a partially stochastic or a wholly stochastic design is most preferable for cost-effectiveness analysis?
(3) Are piggy-back economic studies likely to be appropriately powerful to avoid type II error (i.e. not detecting a *true* difference)?
(4) Can the power be increased by increasing the sample size and/or by decreasing the variance? Or should we accept a different level of statistical significance for the economic analysis?

(5) Are there possibilities of reducing the random component of variation by identifying disease specific cost drivers, using standard unit costs (pooled), or prespecifying outliers.
(6) Is there a possibility of reducing the systematic component of variation by statistical control of economic risk or by analysing practice patterns.
(7) If individual resource items are tested, how should statistically significant differences in opposite directions be treated?

Discussion

A major question to address is the fact that economists, with the available data, will usually not be able to answer even the most simple questions with the degree of rigour demanded by statisticians. The issue then is whether this means that the way trials are currently performed should be changed, or whether economics should have different, i.e. lower, standards for deciding on the statistical significance of a difference than the clinical standards, e.g. by accepting higher limits of significance level or a lower level of power?

In terms of confidence intervals and determining differences on the basis of group data, there is a need to know a lot more about how to reduce variance at the patient level. Basically, the best thing to do is to build in good baseline data, but it is often not possible to find such natural history data. If it is not possible to come up with prior hypotheses about the expected impact on resource use, each single resource item will have to be costed and tested statistically. By regularly inspecting current practice, a better insight into what the cost drivers are, may be gained and this may lead to a more focused selection of data and to the formulation of more accurate and informative a priori hypotheses for new trials. It is necessary to get away from the concept of doing just one, supposedly final, economic study for each new intervention. It would probably be better to adopt the kind of testing programmes known from clinical trials. A preliminary economic study in phase II and III, with all its limitations, restricting its relevance to the question 'can this intervention be efficient?', would then be used to formulate economic hypotheses concerning the process-specific cost drivers for a more definitive phase IV study.

Another question is whether the habitual statistical significance level of 5% is really necessary for economic data. A predominant attitude was that if the primary endpoint of a trial is clinical, it might be acceptable to assess the economic outcomes with a significance level higher than 5%. A further problem is whether significance tests should be performed for each resource item (keeping in mind, of course, that a certain number (e.g. 5%) will show statistically significant differences purely by chance) or on the total cost, which will be determined by multiplying the resources consumed by their unit costs. Testing for statistically significant differences in individual resource utilisation items will certainly render most studies completely under-powered and will often be without meaning when the treatments compared differ fundamentally in the types of resources required. Testing for each study site individually will also lead to a serious loss of power and may further lead to contradictory results among the sites. Using total costs involves a mixture of deterministic and stochastic data, which do not lead to any additional statistical problems, but the problem of how to aggregate economic data across study sites in a meaningful way has not yet been settled.

What the major cost drivers are and where the data on unit costs for these are to come from should be decided in advance, in the design phase of the trial. It should also be decided in advance whether unit cost data will be pooled from all the sites or whether they will be taken as standard costs from one of the health care systems being assessed. The sources for unit cost data should be specified in the study protocol, but it would of course also be very informative to supplement this with an examination and an analysis of differences when other sources for unit costs are being used.

The sample size required for economic evaluation is usually expected to be larger than that required for the clinical outcomes, because of a presumed higher variability in the economic outcomes. Since the sample size of a clinical trial will generally be determined by considerations other than securing sufficient power to detect statistically significant differences in the economic outcome(s), other means of increasing the power of the economic analysis must be found. These may include the use of multivariate analysis methods, prespecified variables and adding a baseline data collection form to help control variations in practice patterns. The possibility of using survival analysis and Cox proportional hazards modelling to control bias due to censored cost data was briefly mentioned, but not seriously discussed.

CHOICE OF ANALYTICAL PERSPECTIVE AND DECISION CRITERIA

The choice of perspective or viewpoint for an economic evaluation may be the single most important decision in the planning of an analysis, as it determines which costs should be included and how they should be valued. The standard recommendation is to adopt a societal viewpoint, but it is contentious whether cost items such as lost productive activity (whether remunerated or not) due to disease and the costs of future unrelated diseases should be included in the calculations.

Cost-utility analysis is based on the assumption that some measure based on individuals' valuations of their own QoL is an appropriate basis for decisions between treatment alternatives. This individualistic assumption may be questioned, for instance arguing that several other individuals than the patient are affected by the treatment, or that the distribution of health gains among individuals should be given more attention. Even accepting the assumption, it is contentious whether equal weighting of each individual's gain is the most egalitarian principle, and it is also debatable who should make the valuation of health states.

Background material (presented by Erik Nord)

The first issue is whether or not indirect costs should be included in the calculation. This controversy may be solved by clearly distinguishing the positive fact, that indirect effects on production of a medical treatment affect the amount of goods and services available to society for other purposes from the normative question of what weight decision makers should attach to these costs, relative to other types of costs. Economic analysis cannot provide an answer to the normative question, but the magnitude of the effect should be reported.

It is suggested that the concept of societal perspective in an evaluation should be interpreted as the viewpoint of the majority of the population. Accordingly, valuations of health outcomes should be elicited from representative population

samples, ignorant of their own future health care needs. General acceptability of this elicitation procedure would also require that the respondents have a good understanding of the impact on QoL of the various disease states and symptoms and that the procedure should ensure that the values elicited reflect the subjects' own interests as potential future users of the health care system.

Allocating scarce resources in health care essentially means dealing with so-called person trade-offs. In terms of the resources spent on providing health care, this means that, if for instance, treatment A costs twice as much as B, two people can be treated with B for the money it costs to treat one patient with A. Whether or not society wants to spend money on two Bs rather than one A depends on the person trade-off in terms of valuation of outcome.

Conventional approaches to utility measurement are not appropriate to answer this type of question, because they have not been designed to rank programmes in a resource allocation context. Respondents are usually asked to value health states, individually or in relation to each other, but they are not asked to give any thoughts on the implications of their valuation for the overall allocation of resources or for individuals threatened by the disease.

Further, the common perception of QALYs as cardinal measures of individual utility is seriously flawed in two respects. First, it seems difficult to obtain meaningful individual utility assessments of life saving treatments. Faced with a choice between life and death, the individual has nothing to lose by sacrificing everything in order to try to stay alive. Second, the assumption that only maximising the total number of QALYs (and not their distribution) matters has been refuted empirically. To obtain meaningful utility measures that also reflect concerns about equity requires direct person trade-off measurements, where the respondents are asked to compare pairs of health programmes in terms of person trade-offs. One programme may be described as saving one person from dying to a life in good health, while the other may cure x persons in disease state S. The respondents would then be asked to indicate the number x which, in their opinion, would make the two programmes equally worth funding. By this procedure, life saving interventions may be valued in a meaningful way and the respondents may place whatever emphasis they wish on matters such as equity and the severity of the initial state. A number of such exercises, focusing on decisions about future treatment capacity and conducted in various countries, show convergence towards broadly the same valuation of health states. They give high value to life saving compared with life improving procedures, and high value to treatment of severely ill compared with less severely ill. Patients with varying potential for benefiting from treatment receive valuations which do not differ dramatically. The last finding could be interpreted as a preference for equity.

Discussion

There was considerable discussion about the inclusion or not of indirect costs and about their possible measurement and interpretation, but it was agreed that they should only be included in analyses conducted from a societal perspective. One concern though, was the risk of double counting, if indirect costs are included. Double counting may result when outcome is assessed by the (potential) patients' willingness to pay, or with a QoL measure, if the ability to perform one's work is an important factor in the respondents' willingness to pay or QoL.

With respect to the valuation of loss of productive time, some contributors to the discussion argued for the use of an average, national wage rate for all the patients involved, whether they are active on the labour market or not. By this solution, the amount of data to collect and valuations to perform could be reduced and it would be less biased than the usual human capital approach in favour of high income earners at the expense of people not currently on the labour market. The opposing viewpoint was that the analyst should give as detailed a description as possible of the impact of the intervention, without being restrained by fear of the possible (mis)uses of this description.

The discussion did not reveal any consensus about the background paper's central criticism of the use of QALYs for societal resource allocation decisions. One criticism was that the new concept, although perhaps useful for decision making, could be criticised along some of the same lines as QALYs are being criticised in the background paper. For instance, who are the right persons to make the valuation? Related to this is the question of the proper decision context, i.e. whether the respondents should imagine themselves as prioritising among current patients or if they should rather take into consideration their own interests as potential future patients. Some experiences, indicating that respondents may be changing their position on these valuation issues when changing roles, were referred, but the implications of such observations did not become clear during the discussion.

RULES FOR GOOD MODELLING IN ECONOMIC EVALUATION

Modelling techniques are widely used in economic evaluations, ranging from simple extrapolations of empirical data to complex and intricate models, where patients move between a succession of progressive states with transition probabilities depending on actions taken and on chance events. Frequently, the models used in published studies appear like 'black boxes' and there seems to be a growing skepticism against the use of models, by, for instance, regulatory bodies. However, some modelling will usually remain a necessity in economic evaluations, because all the relevant data will not always be available from a single trial.

Background material (presented by Bruce Hillner)

Models can never replace prospectively collected data from randomised clinical trials, but they may be complementary in a number of ways. For instance, by extending the findings of a trial(s) to populations different from the trial sample, and in making inferences about how trial results (efficacy) may translate into actual practice (effectiveness) after the trial. A model may be especially helpful in analysing clinical decision problems if multiple outcomes are important, if there are multiple intervening steps between the intervention and the final outcome, and if short- and long-term toxicities can occur. Models may also come to play a role in the design of trials and help to focus the prospective data collection on those items of resource utilisation or outcomes that differ most between the treatment groups.

The essential steps in constructing a model follow from a delineation of the clinical starting point and identification of the outcomes of interest and the relevant time horizon.

Actions within the control of the decision maker and probable events beyond his control must be identified and ordered structurally and chronologically and numerical values assigned to the probabilities of events and to the valuation of outcomes. Some of the necessary values will typically be taken from previous clinical trials (e.g. frequency of events). while others, like cost estimates, will often have to be assembled from several sources.

The primary reason for the lack of agreement on how to define a good model or how to assess the quality of any actual model is that model building demands a delicate balancing of complexity and interpretability. Some basic requirements for an appraisal are that actions should be shown as choices, all relevant disease outcomes should be identifiable, tests should include adverse events and main categories of results and consequences of interventions should be dependent on prior events.

A more advanced appraisal requires that all assumptions be clearly stated and that the sources used for finding probabilities are described in detail, also motivating the choices made among various possibilities, if relevant. Likewise, the valuation of outcomes like QoL adjusted survival or costs should be described in detail, and the treatment alternatives should be valued according to identical principles. It should also be considered whether the time horizon analysed was appropriate, and whether costs and effects were discounted, if the model considers events over 2 years or more after the initial decision. Sufficient detail on all these issues should be reported so that the baseline results can be replicated by any reader. It should also be asked whether the complexity of the endpoints chosen is appropriate for informing the type of decision problem posed in the analysis.

Discussion

There was general agreement that validation of models for decision analysis have been given inadequate attention so far. For preliminary and exploratory models, careful validation seems to be less important than for models aimed for publication, which may even be intended to inform real world decision making. In these cases, validation is essential. Validation should be a continuous process, where the impact of all new relevant data must be assessed.

It was suggested that the concepts face and structural validity could be useful for a sort of conceptual validation. Face validity concerns whether the model gives an appropriate summary of the possible outcomes and structural validity concerns the logical and chronological sequence of actions, events and outcomes. For some contributors to the discussion, structural validity was a point of central concern and potential conflict between different world views and assumptions. For others, the most important problem with models is rather the lack of relevant data from randomised clinical trials for the various parameters of the model, e.g. probabilities and time to events.

Modelling can never fill the gap created by many years of not doing (relevant) trials in a given field, but an important potential function of models is to contribute to trial design. Preliminary modelling may make data collection more focused by pretesting the interesting hypotheses and identifying the variables of most importance. Modelling may also be used for determination of the relevant time horizon and for threshold analyses in connection with sample size calculations.

CAN THE GAP BETWEEN PSYCHOMETRIC QoL OUTCOME MEASURES AND UTILITIES BE BRIDGED?

Approaches to the measurement of QoL and valuations of outcomes have their origin in different scientific disciplines. The fundamental difference between the concepts, methods and applications of the measurement methods imply that the approaches can complement, but not substitute each other. The psychometric approach is mainly used by QoL researchers and the utility or preference-based approach is mainly used by health economists. The aim of this joint session was to let both groups of researchers express and clarify the extent of the differences from their point of view and subsequently to elaborate on the possibilities of bridging the gap between the two approaches.

Background (presented by Dennis Revicki)

Despite increasing consensus about the conceptualisation of QoL, alternatives exist for the measurement of QoL. Two main approaches are used to evaluate generic health status outcomes, preference-based and psychometric-based health status measurement. Developers of different measurement approaches have failed to show any clear superiority in evaluating medical treatments. There are trade-offs in the assessment of QoL and no single approach can fit the objectives of all studies. Utilities are numbers that represent the strength of an individual's preferences for different health outcomes under conditions of uncertainty. Preferences are the values people assign to different health outcomes when uncertainty is not a condition of measurement. These numbers reflect a person's level of subjective satisfaction distress or desirability associated with different health conditions. The preference-based approach uses one or more scaling methods to assign numerical values (utilities or preferences) on a scale from 0 (anchored as death or worst imaginable health) to 1 (anchored as complete health or best imaginable health). Utility scores represent preferences for health states and allow morbidity and mortality improvements to be combined into a single weighted measure. Standard gamble procedures are used to generate utilities, while preferences are measured by using time trade-off, categorical rating scales and other techniques.

Preference-based measures have some advantages compared with the psychometric-based measures. They incorporate time and risk preferences for different health state outcomes into the measurement process and the scores are easily incorporated into economic analyses. There is, however, some controversy regarding the definition of utilities/preferences and the methods used to derive these scores. Utilities/preferences for some health states vary widely among individuals and may be due to framing effects labelling effects, content and duration of health states, cognitive complexity of the measurement task, population and contextual effects and the different scaling methods. Utilities and preferences, compared with psychometric measures, may not be as sensitive to relatively small yet clinically meaningful changes. Preference-based measures sometimes fail to detect differences that are captured by clinical indicators or psychometric health status measures. However, advocates of preference-based measures argue that differences that cannot be perceived by human judges may be too small to be clinically meaningful.

Psychometric approaches to measuring QoL focus on the presence, frequency or intensity of symptoms, behaviours, capabilities or feelings. Responses to individual questions are

aggregated to create individual homogeneous scales (e.g. physical function, mental health) or summary scales. Health status measures have been used to discriminate among individuals with different chronic diseases at a single point in time, predict future health outcomes and to measure change over time.

Only a limited number of studies have incorporated both psychometric health status and preference-based assessments in evaluations of medical treatments. A Medline search of such studies, covering the period 1985–1995 indicates that psychometric-based health status scales are poorly to moderately correlated with standard gamble and time trade-off scores. When combinations of different health status scales are used to predict utilities in regression models, explained variance ranges from 18 to 43%. For rating scale preferences, 27–55% of the variation can be explained by measures of functioning and well being. Rating scale preferences are more closely correlated with various individual health status measures.

In general, there are stronger correlations between a number of different measures of health status and quality of well-being scores. However, measures of psychological well-being tend to have a weaker relationship with quality of well-being scores compared with measures of physical function. These findings may be, in part, due to overlap between the content of the quality of well-being and these health status scales.

The low to moderate association between preference-based and psychometric-based QoL can be explained by the way risk and time is introduced into the assessment process, aspects of the measurement task and the cognitive evaluation processes involved in the standard gamble and time trade-offs techniques.

Preference-based measures and psychometric-based measures are constructed to address different purposes. The psychometric-based measures are designed to arrange persons along different continua of functioning and well-being. The general purpose is to discriminate levels of health status between groups and to detect changes in health status over time. The preference-based scales are designed for application in cost-effectiveness analyses and to help decisions about resource allocation. A single metric on a scale from death to complete health is useful for making judgements about the impact of different health care interventions on health outcomes. Both measurement approaches are useful in evaluating health interventions and are not interchangeable measures of QoL, since they measure different components of health. Investigators need to be aware of the differences between preference-based and psychometric-based QoL measures. Patient assessments of health outcomes are important for the evaluation of medical treatment, and multiple methods for assessing health outcomes are needed to help clinicians, patients and their families in selecting among alternative medical treatments. Future research needs to focus on explaining measurement and individual variation in preferences, and to improve identification of the cognitive processes people use to make preference judgements. Combining health status and preference measures with traditional indicators of safety and clinical efficacy provides more comprehensive information on the impact of medical treatments on patient outcomes.

Discussion

The first issue discussed extensively was the serious concern about the low correlation coefficients observed in several studies. In many studies, the investigation of a relationship between the two measurement approaches was not the primary objective and the design of these studies may accordingly not have been appropriate. The fact that only low to moderate correlations have been found may, to a certain extent, be attributed to the study populations. For instance, the use of homogeneous groups of patients, as required in clinical trials, might result in a low variance of scores and, thus, result in low correlation coefficients. Conducting studies in large samples of the general population implies that 95% of the subjects are healthy individuals who would not be willing to trade-off any life time as measured by the formal utility measurement approaches.

This ceiling effect would lead to low correlations with the results of the psychometric approaches. Another possible bias is the order effect, when multiple measures are used. It is not sufficiently clear to what extent scores are influenced by asking subjects related aspects in distinct ways. Despite the uncertainties about the type and amount of evidence produced so far, it was generally agreed that the psychometric and utility approaches of measuring (health-related) QoL do not seem to be interchangeable. The second issue of discussion was the meaning and consequences of low to moderate correlations. The problem is that with both the psychometric and the utility measures, no gold standard (either internal or external) is available. There seemed to be agreement among both groups of researchers that there is insufficient knowledge on the relative performance characteristics and the underlying constructs that both types of instruments are measuring. As a consequence it is difficult to know if meaningful comparisons are being made.

The next issue discussed was that, if there are conceptual differences underlying the two approaches, removing the gap would not only be impossible, it would not even be desirable. In that case, the solution is to use both approaches simultaneously in order not to lose information. However, from a practical point of view this solution may often not be feasible. It might, therefore, be desirable to use the data obtained using one approach to predict or estimate the outcomes that would be obtained with the other approach, but there seemed to be consensus that it was not yet appropriate for such an endeavour because the 'bridge is too shaky'.

The last main point of the discussion was the relationship between the various methods to measure utilities. From the literature, it is evident that different methods yield different utility values. This seems mainly due to the difference in cognitive tasks required of the respondents. The visual analogue scale and the rating scale are choiceless valuations of outcomes, resulting in scores that tend to be systematically lower than those obtained by means of methods that involve a choice or a trade-off such as the standard gamble and the time trade-off methods. The visual analogue scale and the rating scale methods also seem to correlate better with the overall QoL scores obtained by psychometric methods.

HOW TO PERFORM QoL AND ECONOMIC EVALUATIONS IN LARGE, SIMPLE TRIALS?

Studies aiming at distinguishing the moderate gains realistically anticipated from most new treatments must be designed to avoid both moderate biases and moderate random errors and this implies a need for large numbers of properly randomised patients. Precisely when the expected gains in survival are modest, costs and QoL issues will

probably be assigned greater importance than otherwise. However, QoL assessment and measurement of treatment cost will necessarily tend to make trials more complicated.

Background material (presented by Kevin Schulman)

The first issue concerned some problems inherent in conducting multinational trials. There are great differences between countries in standard clinical practice and in cultural perceptions. Unless the analysis is stratified by country, the analyst may not be able to say anything specific about the individual country.

Some knowledge about practice pattern variations and health system organisation in the various countries is necessary for designing such studies. The commonly occurring variations provide a sort of natural experiment and they should be carefully studied in order to stimulate the formulation of new hypotheses. Further, conducting a pilot study in a single country may lead to an incorrect estimation of the variance of the data, but such an estimate is essential for sample size calculation.

The second point is the question of including economic and QoL assessments in clinical trials and the consequent extra demands on data collection. The question is not only how much room there will be for collection of economic data, but also how this should be approached. For instance, should a minimum set of data be collected for everybody or should specific sub-studies on sub-samples of patients be conducted, and how should this be organised in practice?

A further point in relation to conducting multinational clinical trials is that the discussion so far has focused on the linguistic translation of measurement instruments at the expense of largely glossing over their interpretation. The advantage of a utility measure is that you do not have to know what it means, it means whatever it means to the respondent, whereas a multi-attribute scale or a health status profile means different things in different cultures. Finally, the ethics of resource allocation decisions should also be discussed and assessed, especially when ranges of treatment variations are observed.

Discussion

Firstly, the very idea of conducting large and simple trials was questioned. Trials should be appropriately sized and of adequate complexity and it is questionable whether large trials are needed for assessment of QoL aspects or economic aspects. The basis for randomisation is genuine uncertainty in the clinicians' minds over the treatments' relative efficacy and to recruit more patients than required to answer this question would be considered unethical.

Neither was the concept of a 'large, simple trial' self-evident to all participants and there were different interpretations. For instance: 'if a trial is carried out as a multicentre trial, it can no longer be considered as simple'. To others, simple refers to what one expects to achieve from a trial and from simple trials the decision makers want a final outcome and a cost figure, like an effectiveness trial, looking at a decision to treat in actual practice.

This discussion turned into the question of when to collect resource utilisation data in a randomised clinical trial and how to decide which data to collect. Some participants stated that whenever there is a clinical question, there is an economic question of interest. Others took the position that cost data should only be collected in cases where there is an anticipation of major differences in the net costs of the interventions or when serious rationing of the experimental treatment may be expected. Clinical considerations should be paramount in all other cases.

With respect to adaptation of QoL measurement instruments, the tendency in the past was for investigators to adhere exactly to the original instrument in their translation. The currently favoured alternative is rather to adapt the instruments to the cultures in which they will be used. Such adaptations must be done skillfully, if the possibilities of aggregation across countries are not to be compromised. However, some made a plea for focusing the limited research funds available on determining (clinically) meaningful differences in outcomes and in costs, instead of using a lot of resources on (re)validation of the various existing measurement instruments.

BY WAY OF CONCLUSION

During the sessions of the workshop, probably all the important methodological issues in carrying out economic and QoL evaluations in connection with (cancer) clinical trials were examined and discussed more or less exhaustively, in often very lively debates. The structure and format of the workshop was not intended to result in the formulation of a set of clear guidelines for the proper conduct of such evaluations, and there were no attempts to arrive at a consensus on specific issues. Despite this, a core of agreement was discernible in most of the discussions.

The general impression was that, whilst it is at present not possible to draw any final conclusions concerning the contentious methodological and statistical issues in both fields of research, this does not mean that just about anything goes. There is a limited set of interesting and pertinent solutions, which should be further explored in future methodological research in both fields.

With respect to the field of QoL research, it became clear that some of the problems will be solved in time when more data become available (e.g. the problem of the definition of clinically important changes and sample size estimation) and that a pragmatic approach is currently the best option. However, there are also problems that are more fundamentally related to the concept that is assessed, e.g. the problems of missing data, a single outcome measure from a multidimensional construct, and the integration of length of life and QoL data. To solve these problems, more creativity and testing of existing sophisticated techniques will be required. With respect to the field of economic evaluation, a central concern seemed to be the appropriateness and relevance to policy making of integrating these evaluations in clinical trials, either completely from the early planning phase of trials or later on using a 'piggy-back' approach. This was extensively discussed in more than one session. The central divergence in viewpoint mainly concerned the timing of the analysis rather than the interpretation and significance of its results.

For policy purposes, e.g. in making resource allocation decisions in a single hospital or in a national health care system, it is essential that the evaluations used to inform such decisions are truly comparable. This consideration definitely points to the requirement for adherence to certain methodological norms, but it is not yet clear how restrictive these should be.

A decisive argument in favour of carrying out economic and QoL evaluations in connection with clinical trials is that

Table 2. Ten points for good practice in economic evaluation in healthcare

1. Was a well-defined question posed in answerable form?
2. Was a comprehensive description of the competing alternatives given (i.e. can you tell who? did what? to whom? where? how often?)?
3. Was there evidence that the programme's effectiveness had been established?
4. Were all the important and relevant costs and consequences for each alternative established?
5. Were costs and consequences measured accurately in appropriate physical units (e.g. hours of nursing time, number of physician visits, lost work days, gained life years)?
6. Were costs and consequences valued credibly?
7. Were costs and consequences adjusted for differential timing?
8. Was an incremental analysis of costs and consequences of alternatives performed?
9. Was a sensitivity analysis performed?
10. Did the presentation and discussion of study results include all issues of concern to users?

many decisions for the general adoption of new therapies (e.g. about the reimbursement status of new pharmaceuticals) are taken based on the evidence of the results of phase III studies. However, once such decisions are made, the new therapy may be used by clinicians under circumstances that differ significantly from the trial conditions, for instance on groups of patients that would not have been eligible for the clinical trial. The issue to be addressed by the decision makers at this point in time is whether the information on which the decision on reimbursement status was originally taken is still valid. If the answer is most likely no, the subsequent task is to investigate to what extent there is new information available and, whether this new information would change the original decision. New information for reassessment of decisions about use or reimbursement of a therapy may come from several kinds of sources, for instance phase IV clinical studies, observational data and others.

The perceived need for reassessment may arise because the original efficacy data, cost data or both are not considered to be sufficient or adequate. A similar need will arise if one wants to derive relevant information for decisions from economic evaluations based on clinical trials performed in other health care systems, where the resource utilisation may differ significantly, for instance because of differences in the available resources, in their relative prices or other factors. However, under all these conditions, it seems most likely that modelling will remain an indispensable tool for economic evaluations, as it is unlikely that a randomised clinical trial will be performed for each question one may wish to ask concerning the clinical use of a specific therapy. If strict adherence to the basic principle of making each detail in the working mechanisms of such models and the derivation of data for running them entirely explicit is observed, their usefulness as a means of organising thought and discussions will be enhanced. Under such (perhaps idealised) conditions, models may contribute to better informed decisions, even if they are, as matter of principle, generally not amenable to being fully validated.

It is our contention that all the important topics related to the conduct of QoL and economic evaluations of health care interventions in connection with randomised clinical trials were discussed during the 2 day workshop. Not all the issues raised are specifically related to the integration (or not) of such evaluations in the rather artificial environment of randomised clinical trials, dominated by a culture of rigorous statistical analysis requirements and prespecified analysis plans. These topics (e.g. the appropriate respondents for valuing health outcomes) are pertinent for any evaluation, no matter how the data have been collected.

Some of the general issues involved in QoL and economic evaluations were thus singled out for discussion, while other aspects, not necessarily controversial, were left aside. A single example could be the question about discounting both costs and outcomes, i.e. taking into account the differential timing of costs and outcomes and what rate(s) of discount to apply. The economic hypothesis of time reference states that individuals attach greater importance to events (whether costs or effects on health) in the near future than to events in the farther future. The latter should therefore carry less weight in decision making than the former.

An 'industry' of developing guidelines for economic evaluations of health care interventions has evolved in the 1990s, following the lead of the decision by the Australian government to require that requests for public reimbursement of new pharmaceuticals must be accompanied by submission of the results of an economic evaluation of the new drug. The governments of many other countries are considering following this lead and this has resulted in many activities aimed at setting up such guidelines and at rendering individual studies more comparable. Maynard [22] has accused much of this activity for being a waste of resources on reinventing the wheel, with reference to the checklist of questions with which to assess the quality of economic studies set up by Williams 25 years ago [23]. This list (or slight variants of it) has been widely cited and frequently used, e.g. for periodic reviews of the quality of published studies and it commands wide consensus, which is certainly not the same as being adhered to in practice, as witnessed by the frequent, critical assessments of published studies, maybe because of its broad and unspecific formulations. Table 2 (as reformulated by Drummond and colleagues, as cited by Maynard [22]) indicates the general principles that most practitioners and methodologists in the field subscribe to.

1. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The sickness impact profile: development and final revision of a health status measure. *Med Care* 1981, **19**, 787–805.
2. Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. *J Epidemiol Comm Health* 1980, **34**, 281–286.
3. Stewart AL, Hays R, Ware JE. The MOS short-form general health survey. Reliability and validity in a patient population. *Med Care* 1988, **26**, 724–732.
4. Sackett DL, Chambers LW, MacPherson AS, Goldsmith CH, MacAuley RG. The development and application of indices of health: general methods and summary of results. *Am J Public Health* 1977, **67**, 423–428.
5. 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.
6. Hutton J, Brown R, Borowitz M, Abrams K, Rothman M, Shakespeare S. A new decision model for cost-utility comparisons of chemotherapy in recurrent metastatic breast cancer. *Pharmaco-Economics* 1996, **9**, 8–22.
7. Gold M, Russel L, Siegel J, Weinstein M. *Cost-effectiveness in Health and Medicine*. New York, Oxford University Press, 1996.
8. Drummond MF, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes*. New York, Oxford University Press, 1987.
9. O'Brien B, Drummond MF, Labelle R, Willan A. In search of power and significance: issues in the design and analysis of stochastic economic appraisals. *Med Care* 1994, **132**, 150–163.
10. Koopmanschaap M, Rutten FFH. Indirect costs in economic studies: confronting the confusion. *PharmacoEconomics* 1993, **4**, 446–454.
11. Luce BR, Simpson K. *Methods of Cost-effectiveness Analysis: Areas of Consensus and Debate*. Washington DC, Batelle (MEDTAP) Research Center, 1992.
12. Drummond MF, Brandt A, Luce BR, *et al.* Standardizing economic evaluations in health care: practice, problems and potential. *Int J Technol Assess Health Care* 1993, **9**, 26–36.
13. Rittenhouse B. Potential inconsistencies between cost-effectiveness and cost-utility analyses: an upstairs/downstairs socio-economic distinction. *Int J Technol Assess Health Care* 1995, **11**, 265–276.
14. Drummond MF, Torrance GW, Mason J. Cost-effectiveness league tables: more harm than good? *Soc Sci Med* 1993, **37**, 33–40.
15. Karlsson G, Johannesson M. The decision rules of cost-effectiveness analysis. *PharmacoEconomics* 1996, **9**, 113–120.
16. Rittenhouse B. The relevance of searching for effects under a clinical-trial lamppost: a key issue. *Med Decis Making* 1995, **15**, 348–357.
17. Rittenhouse B, O'Brien B. Threats to the validity of pharmacoeconomic analyses based on clinical trial data. In Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials*. Philadelphia, Lippincott-Raven, 1996, 1215–1224.
18. Rutten-van-Mölken MPMH, van Doorslaer EKA, van Vliet RCJA. Statistical analysis of cost outcomes in a randomized controlled clinical trial. *Health Econ* 1994, **3**, 333–345.
19. van Hout BA, Al MJ, Gordon GS, *et al.* Costs, effects and C/E ratios alongside a clinical trial. *Health Econ* 1994, **3**, 309–319.
20. Siegel C, Laska E, Meisner M. Statistical methods for cost-effectiveness analysis. *Controlled Clin Trials* 1996, **17**, 387–406.
21. Fenn P, McGuire A, Phillips V, Backhouse M, Jones D. The analysis of censored treatment cost data in economic evaluation. *Med Care* 1995, **33**, 851–863.
22. Maynard A. Economic evaluation techniques in health care: reinventing the wheel? *PharmacoEconomics* 1997, **11**, 115–118.
23. Williams A. The costs of benefit approach. *Br Med Bull* 1974, **30**, 252–256.

Acknowledgements—Financial support was received mainly through two EU projects: DG V, SOC 94 201948 05F01 ('Economic evaluation of cancer treatment in Europe: exchange of research experience and education for practitioners in oncology') and DG V, SOC 94 203301 05F01 ('Evaluation of factors affecting the QoL of cancer patients and training of cancer treatment professionals in QoL issues').

APPENDIX
PARTICIPANTS IN THE SYMPOSIUM

D. Curran, G. Hoctin-Boes, W. Kiebert, N. Neymark, R. Sylvester, K. Torfs, C. Van Pottelsberghe (EORTC, Brussels, Belgium); N. Aaronson (The Netherlands Cancer Institute, The Netherlands); R. Berzon (Bristol Myers Squibb, Wallingford, U.S.A.); R. Carr-Hill (University of York, U.K.); C. Couvreur (European Commission, Europe Against Cancer Programme, Luxembourg); L. Davies (University of York, U.K.); H. De Haes (AMC, Amsterdam, The Netherlands); D. Dziadziusko (Medical University, Gdansk, Poland); P. Fayers (Medical Research Council, Cambridge, U.K.); R. Gelber, S. Gelber (Dana Farber Cancer Institute, Boston, U.S.A.); M. Groenvold (University of Copenhagen, Denmark); G. Guyatt (McMaster University, Hamilton, Canada); B. Hillner (Medical College of Virginia, Richmond, U.S.A.); J. Hutton (Batelle Institute, London, U.K.); K. Kesteloot (Centre for Health Services Research, Leuven, Belgium); P. Kind (University of York, U.K.); D. Machin (Medical Research Council, Cambridge, U.K.); G. Macquart-Moulin (INSERM U379, Marseille, France); E. Nord (National Institute of Public Health, Oslo, Norway); M. Olschewski (Insitut fur Medizinische Biometrik & Informatik, Freiburg, Germany); D. Osoba (British Columbia Cancer Agency, Vancouver, Canada) D. Patrick (MAPI, Lyon, France); I. Rosendahl (Karolinska Hospital, Stockholm, Sweden); D. Revicki (MEDTAP International, Arlington, U.S.A.); S. Schraub (Centre Hospitalier de Besancon, France); K. Schulman (Georgetown University, Washington, U.S.A.); P. Selby (St. James University Hospital, Leeds, U.K.); K. Simpson (University of North Carolina, Chapel Hill, U.S.A); H. Sintonen (University of Kuopio, Finland); B. Standaert (AMGEN, Brussels, Belgium); A. Stiggelbout (University Hospital, Leiden, The Netherlands);C. Uyl-De Groot (Erasmus University, Rotterdam, The Netherlands); J. Van Busschbach (Erasmus University, Rotterdam, The Netherlands); W. Weng (Integrated Therapeutics Group Inc. Kellingworth, New Jersey, U.S.A.); T. Whelan (Hamilton Regional Cancer Centre, Hamilton, Canada).