Drug Information Journal, Vol. 36, pp. 209–238, 2002
Printed in the USA.
0092-8615/2002

PATIENT-REPORTED OUTCOMES: THE EXAMPLE OF HEALTH-RELATED QUALITY OF LIFE—A EUROPEAN GUIDANCE DOCUMENT FOR THE IMPROVED INTEGRATION OF HEALTH-RELATED QUALITY OF LIFE ASSESSMENT IN THE DRUG REGULATORY PROCESS

OLIVIER CHASSANY, MD, PHD

Délégation Régionale à la Recherche Clinique Hôpital Saint-Louis, Paris, France

PIERRE SAGNIER, MD, MPH

Bayer, Medical Affairs, Health Economics and Outcomes Research, Buckinghamshire, United Kingdom

PATRICK MARQUIS, MD

Mapi Values, Boston, Massachusetts

STEVE FULLERTON

Zynx Health Incorporated, Cedars-Sinai Health System, Beverly Hills, California

NEIL AARONSON, PHD

The Netherlands Cancer Institute, Amsterdam, The Netherlands

FOR THE EUROPEAN REGULATORY ISSUES ON QUALITY OF LIFE ASSESSMENT GROUP*

*Members of the European Regulatory Issues on Quality of Life Assessment (ERIQA) Group: Dr. Neil Aaronson, The Netherlands Cancer Institute, The Netherlands; Dr. Catherine Acquadro, Coordinator of the ERIQA Group, Mapi Research Institute, France; Dr. Giovanni Apolone, Mario Negri Institute, Italy; Dr. Harry Burns, Dalian House, Glasgow, UK; Dr. Olivier Chassany, Hôpital Saint-Louis, France; Dr. Gianfranco De Carli, GlaxoWellcome, Italy; Dr. Dominique Dubois, Janssen Pharmaceutica N.V., Belgium; Steve Fullerton, Cedars-Sinai Health System/Zynx Health Incorporated, US; Dr. Bernard Genesté, Aventis, France; Dr. Asha Hareendran, Pfizer Central Research, UK; Bernard Jambon, Mapi Group, France; Dr. Patrick Marquis, Mapi Values, US; Dr. Pauline McNulty, ICOM Health Economics, Johnson & Johnson, US; Dr. Caroline Miltenburger, SCHERING AG, Germany; Dr. Annoesjka Novak, NV Organon, The Netherlands; Dr. Margaret L. Rothman, Janssen Research Foundation, US; Dr. Pierre Philippe Sagnier, Bayer plc, Medical Affairs, UK; Dr. François Schubert, GlaxoWellcome Research & Development, UK; Dr. Soren E. Skovlund, Novo Nordisk, Denmark; Dr. Marianne Sullivan, Göteborg University, Sweden; Dr. Marc Tomas, Outcomes Research BMS, Belgium; Dr. Suzanne Wait, Bristol Myers Squibb, UK; Dr. Ingela Wiklund, AstraZeneca R&D, Sweden; and Dr. G. Rhys Williams, Knoll Pharmaceutical Company, US.

Reprint address: Dr. Olivier Chassany, Direction de la Politique Médicale de l'AP-HP, Délégation Régionale à la Recherche Clinique, Hôpital Saint-Louis, 1 Avenue Claud Vellefaux, 75475 Paris Cedex, 10, France. E-mail: olivier.chassany@sls.ap-hop-paris.fr.

The added value for assessing the health-related quality of life (HRQOL) in chronic conditions is now well documented for evaluation of treatment effectiveness in clinical trials and as a criterion for licensing new medications and in policy decisions. However, European standards still need to be developed for the measurement and reporting of HRQOL in clinical trials. This is one of the objectives of the European Regulatory Issues on Quality of Life Assessment (ERIQA) Working Group. This document reviews the major issues arising from the selection of an HRQOL instrument; the integration of HRQOL assessment into the research protocol (methodological design, practicalities of HRQOL administration and collection, prevention and handling of missing data); the statistical analysis plan; and the presentation and interpretation of the results. Finally, to gain wider acceptance, whether HRQOL is considered as a primary or secondary endpoint, the scientific principles of clinical trial design should apply to HRQOL.

Key Words: Health-related quality of life (HRQOL); Randomized clinical trials; Patient-reported outcomes; Missing data; Guidelines

INTRODUCTION

THE NUMBER OF CLINICAL trials incorporating measurement of HRQOL (Of the various terms than can be used to depict a patient's perception of health, throughout this document, we have used the term HRQOL.) has increased substantially in recent years and although regulatory authorities do not require HRQOL evidence for approval of new drugs, pharmaceutical companies are increasingly including this evidence as part of their submissions for drug approval. There are a few conditions, such as cancer, for which the Food and Drug Administration (FDA) has recommended the inclusion of HRQOL assessment (1). Patients are essential as the first and foremost source of information. Hence, the need to assess the impact of treatment on subjective aspects of HRQOL in chronic conditions is increasingly acknowledged by clinicians, pharmaceutical firms, and regulatory authorities for evaluation of treatment effectiveness in clinical trials and as a criterion for licensing new medications and in policy decisions. Moreover, measuring the impact of HRQOL treatment may be pertinent in some acute conditions (eg, infectious diseases).

This recognition comes from several major facts:

- When impairments are incurable, or insufficiently understood, and the realistic goal of care is to make patients' lives as comfortable, functional, and satisfying as possible, outcome management primarily concerns HRQOL. Logically, patients are often the only experts on the various impacts of their diseases or injuries, and so they are the ultimate judge of what treatment has achieved,
- Ratings from patients about their experiences of disease or treatment frequently differ in quality and magnitude from those of physicians and health care professionals or from laboratory tests and other surrogate clinical measures. The direction of this discrepancy may not be predicted a priori,
- Research over the last 25 years into the development and validation of HRQOL instruments has created an environment for growing acceptance of these instruments as conceptually and technically valid measures of health outcomes. These patient-based assessments complement, rather than contradict, the traditional biological and clinical endpoints alongside which they are used, and
- The broad HRQOL concept is made concrete by defining core health-related dimensions to reflect functional limitations and well-being along a continuum from condition-specific to general aspects of physical and mental health.

Guidelines have been recommended for the validation of HRQOL instruments (2,3), and for the reporting of HRQOL results from a clinical trial (4,5). Recommendations regarding the evaluation of HRQOL data for

labeling or promotional claims for the FDA have been published (6,7), including specific recommendations in cancer (1). However, European standards still need to be developed for the measurement and reporting of HRQOL in clinical trials. In Europe, except for growth hormone, no drug has yet been approved with a claim of HRQOL benefit.

If advances are to be made, and wider acceptance gained, there are a number of issues related to the collection, analysis, and reporting of HRQOL data in the clinical trial context that require attention. Several authors have reported that past assessment of HRQOL was often flawed (7–11):

- **HRQOL assessment often fails to be well planned in the protocol,** is poorly implemented and followed during the trial, and is inadequately analyzed, reported, or discussed in the study report and in subsequent publications. This is not true in some fields, such as oncology, where there is greater experience of HRQOL assessment (12). However, it remains the case that even in well-designed oncology trials, approval for labeling or promotional claims regarding HRQOL has not yet been given to any anticancer drugs by European agencies,
- **Technical/methodological questions arise related to the planning and collection of HRQOL data.** HRQOL assessment often appears to be added as an afterthought to a clinical trial, without any data quality procedures. The prevention of missing data needs to be addressed, and the handling of such, if or when it occurs, needs to be defined,
- **Reporting and publication rules are not yet clear.** Due to the unfamiliarity of clinical researchers, editors of medical journals, and regulatory agencies with a field which has been, until recently, concentrated in the social sciences, several aspects of publications relating to HRQOL assessment in clinical trials need to be improved (4,8,9,13), and
- **Interpretation of the results from HRQOL studies is not clear to clinicians and health authorities.** Lack of clarity is due to at least two factors: there are different methods of interpretation and theoretical and practical experience is lacking.

In 1999, five key issues were emphasized by the European Agency for the Evaluation of Medicinal Products (EMEA, Committee for Proprietary Medical Products chairman, Paris, November 1999):

1. What is the added value of HRQOL?
2. Do HRQOL instruments have psychometric properties?
3. Are HRQOL instruments internationally validated?
4. Is the analysis plan adequate? and
5. What is the meaning of observed changes?

The ERIQA project consists of a working group of HRQOL researchers from universities, the pharmaceutical industry, and regulatory authorities with the following objectives:

1. To provide European regulatory authorities with guidance on how to assess the quality of HRQOL studies in clinical trials, and on how to evaluate the validity of HRQOL claims, for appropriate decision making,
2. To convince European regulators that HRQOL is a relevant key outcome, that is, a credible criterion in evaluation of medicines, and
3. To make European regulators feel confident about the quality of HRQOL outcomes.

These objectives were established considering the complexity of the European context of medicinal products and clinical research.

In order to achieve the ERIQA project's first objective, this guidance document reviews the major issues arising from the selection of an HRQOL instrument, the integration of HRQOL assessment into the research protocol, methodological design, practicalities of HRQOL administration and collection, the prevention and handling of missing data, statistical analysis, and the interpreta-

tion and presentation of the results. A checklist summarizing all of these major issues can be found in the Appendix and used as comprehensive tool to review all HRQOL files or dossiers submitted to regulators.

Although primarily intended for reviewers and leaders of regulatory authorities, this document may also be used by sponsors and investigators planning clinical trials and by experts reporting results of clinical trials in study reports or publications. Unless otherwise specified, this document refers primarily to clinical trials whose objective is to show superiority. However, most statements could also apply to equivalence trials and some specific issues related to equivalence trials are defined.

DEFINITION OF HRQOL

Even though there is no definite consensus on the definition of HRQOL, and the term to be used (health status, well-being, quality of life, health-related quality of life), most authors express the same concepts with different words (14,15,16). All definitions derive more or less from the definition of health given by the World Health Organization: "health is a state of complete physical, mental, and social well-being and not merely the absence of disease" (17). Moreover, although some definitions sound accurate, most HRQOL instruments are not able to sample these definitions, for example, the definition given by Patrick et al.: "the value assigned to duration of life as modified by the impairments, functional states, perceptions, and social opportunities that are influenced by disease, injury, treatment, or policy" (18).

Recently, FDA presented HRQOL outcomes as one of the components of patient reported outcomes, an umbrella term for HRQOL, satisfaction with treatment, performance measures for productivity assessment, discomfort, bother, and any other patient-based assessment of health status, well-being, or functional status (19).

Nevertheless, most HRQOL researchers agree upon the following statement: HRQOL is a multidimensional, subjective, and ideally self-administered construct:

- **Multidimensional:** usually composed of the following core set of domains (Throughout this document, the terms 'domain' or 'dimension' refer to what they claim to measure [ie, physical function, anxiety, sleep, etc.] and/or to the construct structure of the questionnaire as revealed by multitrait or factorial analysis. The term 'scale' refers to instruments used to measure those dimensions.): physical function and role functioning; psychological well-being; social functioning; and for specific instruments, disease- and treatment-related symptoms (20). Other domains of HRQOL may be evaluated, depending on the disease and treatment or the condition (general health perception, pain, vitality, sexual function, sleep, etc.),
- **Subjective:** as HRQOL refers to the patient's subjective perception of the impact of disease and treatment on health status, influenced by his experiences, beliefs, and expectations (6,21,22), and
- **Self-administered:** ideally HRQOL assessment should be patient-driven. However, there are situations where a subjective assessment is made via an interviewer (eg, elderly or disabled subjects or subjects with neurological disorders) (20).

What is Not an HRQOL Measure?

There is a tendency to abuse the term HRQOL, and the description of Feinstein is still apposite, "the idea has become an umbrella under which are placed many different indexes dealing with whatever the user wants to focus on" (23).

HRQOL is not:

- A diagnostic scale,
- A disease/symptom severity scale index,
- A symptom listing when the questions ask about the presence or absence of symptoms,
- An adverse event listing (eg, "Have you had a rash in the last month?") (24). Of

course, many HRQOL instruments question the patient about the negative impact of treatment on the different domains of life (25), but the phrasing of the items is very important. An HRQOL questionnaire will ask about the impact of such a symptom on HRQOL, and should never be regarded as a proxy or substitute for traditional adverse event reporting, and
- A single domain.

Although the following domains are part of HRQOL assessment, HRQOL is not correctly sampled with any one alone:

- A disability scale,
- An anxiety or depression scale,
- A fatigue or pain scale, or
- A symptom bother scale.

For example, if in a clinical trial the assessment includes only a functional scale, then providing that it is well validated, the results are comprehensively reported, and the report and interpretation comply with agreed standards, this assessment will be acceptable if the conclusion does not refer to an improvement of HRQOL.

Recommendation

A clinical trial design with HRQOL assumptions should include a conceptual and measurement model to define exactly what is being measured, which domains are covered in a precise condition, and what is the intended HRQOL claim.

ADDED VALUE OF PATIENT-BASED HRQOL ASSESSMENT

The aging of the population in developed countries, and the shift of health care resources from the treatment of acute and/or immediately life-threatening diseases to the treatment of chronic illnesses, has resulted in a need to measure functional status and HRQOL. This need is due to the fact that medical treatment for chronic illnesses is often not curative but aims to improve function by reducing the illness severity or by limiting disease progression.

Traditional symptom, physical, physiological, or biochemical measures of disease activity are the main tools employed to evaluate the pharmacological effects of drug therapy; they do not reflect the patient's perception of his function and well-being. Yet, these perceptions are largely responsible for whether or not a patient considers himself to feel well or to benefit from a treatment (16,21,26). A modest correlation of physiological measures and symptoms with functional capacity and well-being has been reported (27–31). Patients with the same clinical scores (eg, similar range of motion and rating of back pain) may have dramatically different coping strategies and thus their condition has a very different impact on their HRQOL (eg, different role function and emotional well-being scores) (32).

The importance of the impact of disease upon HRQOL may not always be predictable. In a study of the subjective effects of chronic illnesses (back problems, diabetes, and chronic lung problems), patients with moderate gastrointestinal disorders rated the impact of their condition on well-being, mental health, and functional status as worse than that of five other illnesses. Gastrointestinal disorders were matched or exceeded only by the impact of heart disease (33). The burden of asymptomatic mild diseases may be greater than currently perceived by clinicians and sometimes may be similar to that of a symptomatic chronic disease. For example, HRQOL assessment of patients with chronic hepatitis C showed a poorer score than that found in diabetic patients. This reflects, at least in part, the fact that these patients worry about the level of their hepatic enzymes, just as HIV patients are anxious about their lymphocyte CD4 count (34). On the other hand, the HRQOL of surviving breast cancer patients is equivalent to or even higher than scores in the general population (35). In conclusion, the need for, and added-value of, incorporating patient-based assessment and especially patient-driven measures of HRQOL in clinical trials is now well documented; HRQOL data com-

plement well the other data collected during clinical trials.

When is HRQOL a Relevant Endpoint in Clinical Trials?

HRQOL can be useful for therapies that alleviate chronic diseases:

- When one or more HRQOL domain(s) is critical for patients,
- When there is no objective marker of disease activity (eg, migraine, arthritis),
- When a disease can only be characterized by several possible measures of clinical efficacy (eg, asthma),
- When a disease is expressed by many symptoms (eg, irritable bowel syndrome),
- When treatment extends life, possibly at the expense of well-being and HRQOL due to morbidity, functional or psychological impairments, or side effects of therapies,
- When the new treatment is expected to have a small or nonexistent impact on survival (eg, in cancer or chronic conditions) but a positive impact on HRQOL (eg, alleviation of pain, anxiety, stress, or improvement in function),
- With highly efficient treatment in severe and handicapping diseases (eg, rheumatoid arthritis) to ensure that improvement of severity score is accompanied by improvement of HRQOL,
- With not very efficient treatment in less severe diseases (eg, benign prostatic hypertrophy, urinary incontinence) to ensure that the modest improvement of symptoms is accompanied by improvement of HRQOL,
- In diseases with no symptoms (eg, hypertension) to ensure that treatment does not alter HRQOL, and
- In equivalence trials, for drugs anticipated to result in a similar disease course, but expected to have HRQOL differences.

Some acute events (eg, stroke, myocardial infarction, pneumonia, viral infections) also affect HRQOL. In these situations too, HRQOL assessment may be useful except when treatment is very efficient and well tolerated.

SELECTION OF AN HRQOL INSTRUMENT

The selection of an HRQOL instrument is based on the HRQOL hypothesis being tested in the trial. The choice of domains is influenced by the severity and nature of the disease and the expected benefits and side effects of treatment. For example, most patients with severe heart failure are elderly, retired, and physically inactive, so benefits of treatment are likely to be improvements to physical and social functioning, which are already impaired (36).

It is important to select the most relevant, valid, and responsive questionnaire, which should also be available in several languages (in the case of multinational studies). It must be verified that the domains explored by an instrument reflect those which are expected to change (37).

Type of Questionnaire

Generic Versus Specific Questionnaire.

- **Generic instruments** are designed to assess and compare health status among a broad range of patients with different health states, conditions, and diseases (21,36,38). The validity and reliability of these instruments (eg, MOS 36-Item Short-Form Health Survey [SF-36], Sickness Impact Profile, Psychological General Well-Being Index) are generally very well documented as they have been tested in many conditions. As they were developed for general populations and do not focus specifically on the impact of a particular disease, generic measures are less likely to detect small but clinically important changes over time or induced by treatment (21,22,25,36,38–41). Thus, in a study of 424 patients with gastro-oesophageal reflux disease, the Psychological General Well-Being Index, a generic HRQOL instrument, was unable to detect any difference between the HRQOL of pa-

tients receiving omeprazole and those receiving cisapride, even though there was a significantly greater improvement in reflux symptoms in the omeprazole group (42). Only a large improvement is likely to be detected by generic instruments. For example, another study showed that liver transplantation was associated with substantial improvement in life quality as assessed by the Sickness Impact Profile, although as a group, patients undergoing this surgery did not recover to the level of functioning demonstrated by normal individuals (43). The use of a generic instrument may also reduce the chance of detecting an unexpected effect if no item addresses it. For instance, the SF-36 questionnaire contains no item relating to sleep, which is impaired in many conditions (44), and

- **Disease-specific instruments** include items that are likely to be affected by the particular disease or situation under study, and consequently are expected to be more responsive than generic measures to clinical changes. Specific instruments, therefore, are more appropriate for clinical trials designed to evaluate specific treatments (21, 38). Specific questionnaires may also reduce patient burden and increase acceptability as they only include relevant dimensions, unless, of course, the number of items is too high.

Whether a generic or specific questionnaire is used, there is a possibility of missing effects if some domains are not included in the questionnaire (eg, some impact of adverse events).

Battery of Questionnaires. The battery approach involves the use of several instruments. A generic health-status measure may be used together with one or more condition-specific instruments and one or more domain-specific scales (eg, sleep, diet, pain, sexual function, coping with disease). This approach raises several problems that are not only found with batteries, such as complexity and length of the case report form (CRF); burden on patients and investigators; missing data; multiple statistical test comparisons; and the impossibility of having a single score. Moreover, unless one instrument has been selected to be the primary endpoint, the interpretation of results of different instruments used concomitantly in a clinical trial is difficult, particularly if results are contradictory. Another concern is the validity of a battery, as the evidence of validation of each instrument included in the battery does not ensure the validity of the battery as a whole in the target population. Finally, whether the various instruments reflect similar concepts and domains must be checked.

Single HRQOL Item, Single Satisfaction Item, Global Question. A single item of HRQOL such as "Do you consider your health to be excellent, good, fair or poor?," although it is known to be a satisfactory item from a measurement standpoint, does not provide information about the level of specific components or dimensions of health and may be subject to a wide variety of interpretation (45,46,47). Single item scales, irrespective of whether they tap health status, HRQOL, satisfaction or ask about global treatment evaluation, are compromised in terms of reliability and precision of the measurement.

Short Forms of Questionnaires. Short forms of some questionnaires (eg, SF-12 derived from the SF-36) have been developed in order to reduce the burden on the respondent. Even if the structural properties remain identical to the original questionnaire, short forms of questionnaires are less accurate and responsive.

Recommendations

- *The selection of the HRQOL measurement strategy should be justified, including details of the validation process, and should include references in the research protocol, the study report, and any publication(s). This is particularly important if the questionnaire is not well known (37). For a questionnaire such as the SF-36, the valid-*

ity of which is widely recognized, a more succinct description with references is sufficient, and

- *In situations where the benefit in terms of global HRQOL is expected to be rather small, a specific instrument should be used, providing its psychometric properties are good. A generic instrument (sometimes assessed only at baseline) can be used alongside the specific instrument as it captures different domains from the specific questionnaire. Furthermore, its scores will allow comparisons between studies which may help in the interpretation of results (36). In this case, the generic questionnaire completion should precede the specific questionnaire to allow comparisons across studies.*

Instrument Properties

The development of an instrument is a long process, involving item generation; item scaling; item aggregation into domains; item reduction; and demonstration of acceptability, reliability, validity, and responsiveness. This process begins with the conceptual and empirical basis of the measurement model underlying the instrument. The instrument selected for a trial should have demonstrable evidence of reliability and validity in the target population (2,48,49), unless these properties are to be evaluated during the trial.

Item Scaling and Scoring of Instruments.

- Questionnaires consist of items aggregated into different domains (the scale structure is verified by statistical analysis using factorial analysis and by content validity),
- For each item, a patient's answer is requested. To ensure questionnaire responsiveness, answer options must contain sufficient gradations to detect small changes. Many scales include items with response options on a four- to seven-point Likert scale. A linear visual analogue scale more closely approximates true interval level measures but the required level of abstraction may be too great for some patients (20). A simple yes/no answer option is less responsive,
- Usually, the answer to an item, for example, on a five-point response scale from 'none' to 'extremely,' is transformed to a score ranging from zero to four. The scores of the domains are computed by summing the item scores (50). To facilitate the presentation of results, domain scores can be linearly transformed into a range from 0 to 100 (a percentage of the maximum achievable score) (36,51). The direction of change should be defined (eg, 0: worst HRQOL, 100: best HRQOL).
- Depending on the questionnaire, the results will be presented as separate scores for the different domains (profile), and/or as a single score (index) or summary scores (eg, physical and mental scores of the SF-36) which aggregate scores across the domains (36). The advantage of a single score is that it simplifies the statistical comparisons and facilitates interpretation. Its main drawback is that it may obscure information, as a global score may show a statistically significant difference in favor of a treatment, while masking deterioration in an important domain (37,45),
- Item or domain scores of some questionnaires are weighted. Weightings can be derived from statistical procedures or based on preference of patients or experts during the validation process or from the importance given by the patient to each item. Weightings can be included in the standard scoring algorithm, and
- It may be uncertain whether the statistical procedure of weighting which has been selected in one setting is appropriate in another context (52,53). However, asking the patient to weight each item or domain in addition to completing the questionnaire is time consuming during a clinical trial and is likely to result in more missing data (36). Moreover, interpretation of results may be more difficult so weighting is currently not recommended; if used, it should be kept simple (50). Simple uniform scoring schemes, giving the same value to each item, are preferable. An implicit weighting exists

when some domains contain more items than others and when a global score is computed by summing the answers from all items. It seems preferable when computing a global score to sum the scores of the different domains, which then have equal value.

Reliability. The reliability of an instrument is expressed by a coefficient, which may range from zero to one, with one indicating maximum reliability (2,3,54). This coefficient is estimated using the correlation among items (internal consistency), and the test–retest reliability. When the questionnaire is administered by interviewers, the interrater reliability should also be assessed (45). Reliability includes:

- **Internal consistency reliability** assesses the correlation across items within the different domains of the questionnaire and across all items in the questionnaire. Assessment is usually carried out using Cronbach's α-coefficient; a minimum value of 0.70 is usually recommended for group comparisons (3,36,45,48,49,51,55). For instruments using dichotomous response options, the Kuder–Richardson formula may be used (2),
- **Test–retest reliability (reproducibility)** is the degree to which an instrument gives similar scores on repeated administrations in identical conditions to respondents who are assumed to be stable with respect to the domains being assessed. It is often based on the intraclass correlation coefficient or Kappa coefficient (ordinal numbers); a reliability coefficient greater than 0.70 is considered acceptable for group comparisons (2,3,7,45,47). The time frame between the two administrations should be between one and four weeks, depending on the disease (2,3,49). Test-retest reliability should be confirmed in a study conducted in at least 30 stable patients (in parallel with the number of patients needed to perform a parametric statistical test). For example, assessing the reliability of an HRQOL instrument by administrations 1 month apart to 19 patients who had reported no change in their disease state is probably inadequate (44,56), as are repeated administrations 48 hours apart (57), and
- **Interrater reliability** is an estimate of the extent to which two or more trained interviewers scoring the same items simultaneously agree with each another. A correlation of 0.80 or higher between interviewers is desirable (45).

Assessing reliability is not sufficient to validate a questionnaire. A nonresponsive questionnaire (eg, binary response options, or items that measure a personality trait rather than subject status) will exhibit a high reliability coefficient, but will still be inadequate for measuring change over time or difference between treatment groups.

Validity. This is the degree to which an instrument measures what it is intended to measure (2,3,45). Demonstrating validity is a rather difficult goal to achieve, as the measures used to evaluate the correlations within the HRQOL instruments are rarely validated themselves. Validity is based on a conceptual and empirical scheme for gathering items into domains and/or a single scale (2). Practically, it is the extent to which the new HRQOL measure correlates with other HRQOL instruments and clinical measures, and its power to discriminate between groups varying in disease severity. Three types of validity are usually required: content, construct, and criterion validity:

- **Content validity** is the extent to which the domains of interest are comprehensively sampled by items and whether the questionnaire covers the full range of relevant topics. This process involves speaking with experts and patients in order to check the clarity, comprehensiveness, and acceptability of questions and answer options, and the absence of redundancy of items (2), and
- **Construct validity** is assessed using different methods:
  - **Structural validity** is a statistical procedure using factor analysis, multitrait

analysis, or Rash analysis to support the hypothesized scale structure of the questionnaire (ie, the combination of items into dimensions) (2). In factorial analysis, items reflecting the same concept tend to be associated in the same factor. In multitrait analysis, which is based on item–scale correlations, item convergent validity is accepted when an item is correlated with its own domain (ie, Pearson's correlation coefficient ≥ 0.4), and item divergent validity is accepted when the item is more strongly correlated with its own domain than with other domains (32,47,51),

- **Clinical validity** explores logical relations that should exist with other measures (45). It is established through the development of hypotheses about the behavior of scores of the HRQOL instrument in various situations. An HRQOL measure should discriminate between groups of patients whose health status differs, according to the characteristics of their disease, for example, disease severity (so called discriminant or known groups validity) (54). For example, the discriminative capacity of the Functional Digestive Disorders Quality of Life (FDDQL) specific questionnaire was established as the patients with the most severe disease (in terms of handicap, the number of symptoms, and digestive status) reported significantly lower HRQOL domain scores and global scores than the others (44),
- **Criterion validity** measures the correlation between a new instrument and a reference and is difficult to achieve because of the absence of widely accepted 'gold standard' measures. It is approached by correlating the new HRQOL instrument with generic or specific instruments that are considered to be validated and that evaluate similar, related concepts (32, 36,45,51). A new HRQOL questionnaire should only yield a moderately close correlation ($r \approx 0.4–0.7$) with a well-established scale. If that correlation is very close ($r \approx 1$), it means that the new HRQOL tool is redundant (54). For example, the concurrent validity of the FDDQL questionnaire was supported as the scales of the specific FDDQL and generic SF-36 questionnaires exploring the same concepts (ie, physical) were more closely related than scales exploring different concepts, the SF-36 questionnaire being taken as a reference (44), and
- **Predictive validity** explores the correlation with another measure assessed in the future (ie, ability to predict future mortality or morbidity). Several studies have shown such a predictive value for HRQOL questionnaires in diseases such as chronic obstructive pulmonary disease or cancer (58,59,60). For example, a 16 to 30 point decrease in the global score (range: 32 to 224, worst–best) of the Inflammatory Bowel Disease Questionnaire is considered meaningful (61) and when compared with the validated severity index Crohn's Disease Activity Index, it is predictive of a relapse and is also associated with a change of therapy by the practitioner (61). Moreover, the social function domain may be predictive of a need for surgery, the presurgical score being impaired compared with relapsed patients who have been successfully treated by medical treatment (61).

Responsiveness. This is the ability of an instrument to detect small but important changes over time (delta change from baseline in a group of patients) or differences between treatment groups at a specified time (62,63,64). The responsiveness of a new instrument can be compared with that of other instruments by using effect size or standardized response techniques. This property may not have been tested prior to the use of the instrument in a clinical trial. In such cases, responsiveness may be approached by the capacity of the instrument to discriminate between clinically meaningful groups at a single timepoint (ie, known groups validity, see above), but this is not always proof that the instrument will be sensitive to HRQOL change over time or to differences between

treatment groups. Conversely, if the objective of the clinical trial is to show equivalence between drugs, then evidence of responsiveness has to be proven before the start of the trial.

Language Availability. For international trials, the questionnaire selected should be available in different languages. Linguistic translation and validation is a rigorous process designed to ensure that the same meaning of the original concepts exists in all translations (65). The methodology uses at least two independent forward translations, followed by a quality control process involving backward translations, international harmonization, and cognitive debriefing involving lay panels of patients to ensure that items reflect the same concepts as the original scale (66–69). The translated versions must display the same scale structure (construct validity: item aggregation into dimensions) (70) and should have been tested in terms of reliability and validity to enable pooling of HRQOL data from the different countries participating in the trial. The verification of these properties in each country (or in case of small numbers of patients, by pooling languages according to similar cultures, for example, Northern, Latin, and Eastern countries) can be achieved prior to the trial by validation studies in each country or on HRQOL data obtained in each country during the trial.

Acceptability. Elements of acceptability, such as the time needed to complete the questionnaire, the physical and mental ability of the patient population, the rate of refusal of completion, and the percentage of missing items per questionnaire, should be explored before selecting an instrument.

In summary, the major issue in instrument selection is how much validation evidence is enough (6). This justification is critical when the use of a new and unknown instrument is planned. If no validation data exist or are provided, criticism could be raised that the instrument has been purposely designed to focus on the specific drug under investigation (ie, the sponsor decided what it wanted to ask) (7).

Recommendations: Minimal Level of Validation Required

- *The instrument should have demonstrable evidence of reliability; construct validity, including clinical validity; and responsiveness over time in a population similar to the one expected to be included in the clinical trial,*
- *Internal consistency: a minimum value of 0.70 is usually recommended for group comparisons,*
- *Test–retest: a reliability coefficient greater than 0.70 is considered acceptable for group comparisons,*
- *Concurrent validity: a moderate correlation of 0.4 to 0.7 is desirable between the questionnaire under validation and other measures,*
- *The ability of the questionnaire to discriminate between groups of patients differing in severity is of greatest importance,*
- *The validation data will help to calculate the sample size,*
- *If the target population of the trial is different from the patients included in the validation process, the questionnaire should be validated prior to its use in the trial,*
- *Similarly, if the questionnaire is validated in one mode of administration (eg, self-administered questionnaire), it should be validated for use in another mode prior to its use in a clinical trial,*
- *If applicable, the different translations of the questionnaire should have undergone a linguistic validation process, and*
- *It is acceptable that a clinical study also serves simultaneously as a validation study to assess reliability, construct validity, and responsiveness over time. It is essential that statistical analysis for psychometric validation be performed using blinded data before the analysis of results from different treatment groups as part of the clinical study (ie, the item selection and scoring algorithm should not be biased toward the*

treatment effect). In these cases, there is a risk that no change over time or between-group difference will be shown if the questionnaire lacks validity or responsiveness.

STUDY DESIGN— PRACTICAL CONSIDERATIONS

HRQOL claims cannot be based on noncomparative and nonblind clinical trials, which always lead to a higher rate of positive results and for which no causal link can be established between the therapeutic intervention and the HRQOL change (46,71).

The study design must be carefully defined as the best research is always hypothesis driven; HRQOL should not be an add-on to an existing clinical trial. (11). The same rigorous clinical research standards apply as for any physiological or clinical endpoint, regardless of whether HRQOL is considered a primary or secondary endpoint. There are several issues that need to be justified and planned in the protocol regardless of whether HRQOL is a primary or secondary endpoint.

Timing of HRQOL Assessment

- There should be a reasonable number of HRQOL assessments during a trial; the precise number will depend upon the disease and its severity; the risk of early death or loss to follow-up due to adverse events; the expected maximum response to treatment; the length of the trial; and the number of questions to be answered. Assessing HRQOL every week during a two-month trial in patients with a chronic nonlife-threatening condition is of no interest, except if the treatment acts very quickly or if adverse events are likely to appear during the initial part of treatment. In most situations, following a baseline assessment, a sufficient length of time before HRQOL assessment must be allowed for the HRQOL changes to occur; this may be different from the time for clinical changes to appear. Furthermore, time must be allowed for any placebo effect and/or side effect(s) to resolve, if they are going to (25). For example, a patient's HRQOL may improve after being selected for a trial of a new drug for hepatitis C, independent of the impact of the drug on his disease (11),
- Too frequent evaluation may increase the proportion of missing data and type I errors due to multiple statistical tests within the study, unless the statistical analysis plan takes into account these multiple comparisons (6,36,71,72). Conversely, lack of intermediate or long-term follow-up may jeopardize HRQOL results,
- In most trials, a baseline assessment is performed. These baseline values can be used as covariates in the statistical analysis (25, 71,72) and to evaluate responsiveness over time. The baseline assessment can also be used for patient stratification according to HRQOL scores (71),
- The research protocol should specify and justify the number and the timing of HRQOL assessments,
- As a minimum, HRQOL assessment should be performed at baseline and at the end of the study or at withdrawal (72,73), and
- HRQOL assessment is usually performed at the same time as scheduled clinic visits (72).

Standardized Training of Study Personnel

- Before the start of the trial, investigators and their staff and clinical study monitors should be trained to comply with standard operating procedures for the administration and collection of HRQOL questionnaires (72),
- As with any data, these standard operating procedures should describe how the investigators must disclose incomplete data: for example, notes of missed items within a questionnaire and explanations of why a questionnaire was not completed by the patient should be initialed by the investigator, and
- The two points above can be appropriately dealt with by developing an investigators' guide/booklet for HRQOL assessment. This guide should be discussed at length during

the investigators' meeting and serve as a reference document for investigators and monitors during the course of the trial. Another option is to have reminders included in the case report form, which the person responsible for the HRQOL assessments is required to check and sign.

Mode of Administration

- As HRQOL is subjective, assessment is ideally carried out using self-administered questionnaires,
- In some situations, the patient may be unable to complete a questionnaire unaided, and thus assessment through an interviewer or proxy assessment may be useful. It is not well known, however, if the answers given by a physician, a spouse, or another primary caregiver are similar to those which would be given by the patient (20, 74,75) or if they underestimate the impact of the items which most distress the patient (5,76). It has been suggested that patients tend to report a better HRQOL when a questionnaire is interviewer-administered, perhaps because patients want to please the investigator or because they are reluctant to reveal impairment of some intimate aspects of their life (eg, sexual function) (5,25,72,76). It is important to assure patients that the confidentiality of their answers will be respected. In one study evaluating sexual impairment induced by an antihypertensive treatment in male patients, the answers given by patients and their partners were quite different (77). However, recent studies have indicated that both formal and informal caregivers are capable of providing valid and reliable HRQOL data in some conditions such as cancer (78–81),
- The patient should only receive help when it is absolutely necessary, and investigators, nurses, and spouses should all be discouraged from offering help (5). However, assisted completion, when patients are unable to complete the questionnaire by themselves, or proxy completion, when patients are too ill, are better than a total absence of data (5,71),
- Cases of assisted completion and proxy completion should be documented and the reasons given,
- Interviewer-administered questionnaires, including telephone interviews, require extra resources for staff and training to minimize inter- and intra-rater variability, and to minimize verbal and nonverbal sources of bias,
- Telephone interviews are currently infrequent in Europe due to administrative and legal constraints,
- Interviewers and proxies should be consistent during the trial, and
- The mode of administration should be consistent within a single study or, if this is not possible, it should at least be standardized across patients and treatment arms (eg, initial self-assessment in the clinic, with follow-up done either by postal questionnaires or telephone interviewing). If the HRQOL assessment is not clinic-based but is to be done at home, reminder letters and/or telephone calls are needed (72).

Eligibility Criteria

- In most situations, patients who cannot read and write or who are unable to understand how to complete the questionnaires are to be excluded. It may be useful to test that a patient has the ability to read and write and understand the questionnaires in a separate, brief session prior to randomization; the results of these training sessions should not be included in the analyses,
- However, patients with some conditions, such as neurological diseases (stroke, Parkinson's disease, paraplegia, etc.) or visual deficiency, may not be able to complete HRQOL questionnaires without assistance. In these cases, it is acceptable to include them and to allow an interviewer or a proxy to complete the questionnaire, providing the fact that the questionnaire was not self-administered is recorded. This is preferable to no information (82), and
- It may be useful in some situations (eg, nonsevere and nonprogressive diseases such

as irritable bowel syndrome or asthma) to include only patients with a certain level of HRQOL impairment at baseline, in order to include more impaired patients. Another strategy is to stratify the patients according to their baseline level of HRQOL, in order to perform subgroup analyses. However, depending on the results, this may render conclusions difficult or restrictive to make.

Standardization of Data Collection

- If the clinician's visit is likely to be either reassuring or stressful for the patient (eg, giving bad adverse event reports or bad clinical or functional capacity results for a patient with cancer), this is likely to impact the self-perceived HRQOL. In such cases, the questionnaire must be completed before the clinician's visit (5),
- If the trial concerns a nonsevere pathology (eg, functional digestive disorders) and if the visit will potentially not reveal any negative or positive result(s), then completion of the questionnaire may be before or after the visit. Practically, it may be easier to complete the questionnaire after the visit,
- The time of completion must be the same for all the patients during the trial,
- It is important to ensure the patient's privacy and confidentiality (5,36). Therefore, a peaceful place in the clinic should be available during the trial,
- Patients should complete the questionnaire alone, without family or staff present (5, 36), and
- In practice, there must be an interaction between the investigator or study nurse and the patient. Investigational staff play an important role in the acceptability of the HRQOL questionnaire and data quality. Staff must always motivate, but not influence, the patient. Instructions should be given only as described in the manual.

Recall Period

- This is the timeframe that the patient should consider in giving his/her responses to the items of the questionnaire, and
- This period is usually any time up to two to four weeks. People are generally better able to report their experience more accurately over a relatively short time period rather than over longer intervals (20,71). The timeframe should be relevant to the course of the disease or symptoms.

Respondent Burden

- This is the time required to complete the measure and any stress that the respondent may endure as a result of completing the measure (45). A heavy respondent burden may affect both the willingness of the patient to participate in the study and the quality of the data, and
- The number of instruments to complete should be kept to a minimum to reduce the respondent burden and the risk of missing data.

Multicenter Trials

All sites in a multicenter trial must comply with the protocol, and must follow the same procedures concerning administration of the HRQOL instrument and the collection of HRQOL data (5,71). Centers with low completion rates and poor data quality are a concern and may be excluded.

Modification of the Instrument

Instruments should be used in their original validated form, without changing or deleting any part of the original; otherwise, it is not proven that the part of the original instrument being used is valid or if the choice is biased (71). Item scaling and scoring should be as specified by the instrument developers, unless new procedures are made available based upon recent data.

These practical considerations should be planned and described in the research protocol, together with quality control procedures. This will involve extra financial and personnel resources and provision should be made for them in the budget.

Recommendations

- *The research protocol, study report, and any publications should document sufficient methodological detail to permit a critical review,*
- *Randomized controlled clinical trials with double-blind procedures (whenever possible) are a prerequisite,*
- *When a double- or single-blind study is not possible, in the case of an interviewer-administered HRQOL questionnaire, the research protocol must specify that the interviewers have to be blind to the treatment received by patients. In cases of self-administered questionnaires, investigators and other investigational staff must not influence, through verbal or other attitudes, the completion of the questionnaire,*
- *HRQOL assessment of premature withdrawals and dropouts should be performed systematically to reduce missing data (36), and to account for potential selection bias,*
- *Ideally, in some trials dealing with progressive diseases such as cancer, patients who prematurely discontinue their treatment should continue to have their HRQOL assessed until the theoretical end of their participation in the trial,*
- *Informing and motivating investigators is critical (investigators' meetings, monitoring), and*
- *Everyone involved in the trial must apply Good Clinical Practice to the assessment of HRQOL.*

STATISTICAL ISSUES

HRQOL data should be treated like any endpoint and an agreed standard of statistical analysis must be applied (83). Simplicity of analyses should be the rule whenever possible.

Sample Size

HRQOL as a Primary Endpoint.

- An estimation of the sample size for an HRQOL primary endpoint should be performed, as for any endpoint; this is an essential element of the design of any clinical trial (10,50,84). Without prespecified hypotheses, there is a high risk of underpowering the study, leading to a nonsignificant difference between groups; such a study can be considered unethical (85). Conversely, including several hundreds or thousands of patients without any justification in a controlled clinical trial (eg, more than 800 patients in a placebo-controlled trial [86]) is likely to reveal small and statistically significant differences in the scores of HRQOL questionnaires, for which the clinical relevance is difficult to interpret (84,87,88),
- The sample size is easier to estimate when the primary endpoint relies on a single score. In the case of a profile questionnaire or battery, a sample size should be calculated for each of the scores, or preferably for a few scores that have been selected as the most important, then the highest of the calculated sample sizes should be chosen for the trial,
- Estimation of the sample size requires hypotheses about the expected differences in scores between groups, for example, what is the expected effect size (calculated as the difference in score between groups divided by the pooled standard deviation of the baseline scores) or the clinically important difference? The sample size also depends on the level of type I error, which should be adjusted to the number of statistical comparisons, from the power (type II error is generally fixed at 0.10 or 0.20) and the variance of the HRQOL scores (84),
- When comparing two groups with an alpha risk fixed at 0.05 and a power of 0.80, 400 subjects are needed per group to detect a small effect size of 0.20. To detect a moderate effect size of 0.50, approximately 64 subjects are needed per group. To detect a large effect size of 0.80, at least 26 subjects per group would be required (63) (see Table 1), and
- It is of interest to use HRQOL questionnaires in Phase II trials to establish or confirm the validity of potential instruments, to estimate the magnitude of changes among

TABLE 1
Estimation of Sample Size According to Expected Effect Size

Effect Size	Small (0.20)	Moderate (0.50)	Large (0.80)
Number of subjects needed in a two-group parallel trial	800	128	52

the dimensions, and to determine a clinically significant difference. This, in turn, will allow generation of hypotheses for change and a sample size for the Phase III trial(s) (7).

HRQOL as a Secondary Endpoint.

- When HRQOL is a secondary endpoint, as for conventional measures, sample size is rarely estimated. Nevertheless, some hypotheses should be made prior to the start of the trial concerning the expected changes in HRQOL scores (effect size or minimal important difference) given the agreed sample size. For example, if a study includes 2000 patients to show mortality prevention as the primary endpoint, it is likely that a small and statistically significant, but not clinically pertinent, difference will be shown in HRQOL scores. Conversely, a small study may not be able to detect a significant difference in HRQOL scores, and
- If the sample size required for HRQOL assessment is substantially less than for the primary endpoint, an unbiased strategy for selection of a subset of patients in whom HRQOL will be assessed is possible, provided that this strategy is clearly defined and justified in the protocol.

Presentation of Results

Standardized reporting of results is critical for the study report and publication(s) and should include:

- Description of the follow-up of the patients in each group: eligible patients, patients included, patients who completed the trial, minor and major protocol violations, withdrawals (due to: improvement, lack of efficacy, adverse events, unrelated to treatment), and losses to follow-up,
- Procedures for the scoring of items and scales and a description of the possible range of scores and the direction of change (eg, higher score indicates better HRQOL),
- Clear distribution of results between groups (36):
  - mean score and standard deviation at baseline and during the follow-up of all the domain scores in each treatment group,
  - difference between the treatment groups for all scores, with the mean, the standard deviation, the p-value, and the 95% confidence interval, and
  - it is clearly unacceptable to selectively report only positive findings; all findings must be disclosed,
- Description of the statistical tests used, and
- Baseline values can be used as a covariate in the analysis (delta change from baseline in each treatment group or covariance analysis). This method would reduce the influence of baseline differences and make the statistical analysis more efficient; it can also account for some cross-cultural variations.

Intent-to-Treat Analysis

- The only unbiased analysis is an intent-to-treat (ITT) analysis in which all of the patients randomized are included in the analysis population, including patients who did not complete the trial (4). A 'modified ITT' (eg, analysis of patients who have taken at least one tablet of the treatment under study) is acceptable, provided that the percentage of patients excluded from the analysis remains low,

- This analysis may put the treatment under study at a disadvantage, but excluding patients from the analysis, whatever the reason, may result in bias (ie, destruction of the comparability of treatment groups if analysis is only of 'good' patients who continued the treatment) (36),
- Moreover, patients who do not complete the trial are possibly the most interesting with respect to HRQOL (ie, they may have the poorest HRQOL) (20,50,89),
- For example, in a study comparing two antihypertensive agents, nifedipine and atenolol, the change in HRQOL scores from baseline to last available visit showed a much larger deterioration for those who withdrew early from the trial compared with those who completed the study protocol. When treatment differences were calculated based upon all cases (using last observation carried forward [LOCF]) or 24-week completers, the conclusions were quite different. The LOCF analysis showed no treatment difference whereas the completer analysis showed a result in favor of nifedipine (77),
- It is for this reason that patients who drop out should be followed up whenever possible, and their HRQOL assessment should be continued under the same conditions as those patients who completed the trial (25), and
- All analyses should be clearly defined a priori in the research protocol.

Handling of Missing Data

Missing data can be a detrimental to the results of a clinical trial (72,90,91). Like other data, missing HRQOL data (missing items or missing questionnaires) can be related or unrelated to the disease and the treatment. There is strong evidence that missing data are not 'missing at random' (ie, they are related to either the treatment or the underlying disease) and cannot be ignored without introducing bias (6,72). HRQOL data may be missing more frequently than data derived from physiological or clinical endpoints, as HRQOL assessment relies on the completion of many items (72). In an early adjuvant breast cancer trial, HRQOL assessments were terminated as a consequence of poor questionnaire completion rates (92). In practice, patients who do not provide HRQOL data, for whatever reason, may be different from those who do (36).

There are three major issues concerning missing data (prevention, description, and imputation):

Prevention of Missing Data.

- Certain missing data, for example, as a result of death or withdrawal of patients with severe adverse events, are unavoidable. Other missing data can be minimized through standardized procedures described in the protocol,
- The respondent and staff burden resulting from HRQOL questionnaire administration and completion should be reduced to a minimum (ie, the acceptability of the questionnaire selected for the trial must be known before the start of the trial). The use of a battery is likely to increase the risk of missing data, as these questionnaires may need different response formats and, therefore, require multiple instructions and answers to many items (72),
- Clear explanations regarding the rationale for HRQOL assessment should be provided to patients by all investigators,
- Instructions for the completion of the questionnaire should be written on the first sheet of the questionnaire,
- Investigators should understand and agree with the assessment of HRQOL. Questionnaire completion rates in multicenter trials depend more upon the institution and the attitude of the investigator than on the clinical or sociodemographic characteristics of the patient (72,93),
- Investigators and clinical monitors should be trained to administer, collect, and check questionnaires,
- In particular, it is important that investigators and their staff are aware that they are responsible for checking that there are no missing items. Questionnaires are still not

routinely inspected for missing responses after collection (72),
- All patients who withdraw from or drop out of a study must complete, whenever possible, the questionnaire at a time close to when they leave the trial (36). If possible, all patients should be followed until the end of the study, even in the case of a premature end to treatment,
- Too frequent HRQOL assessment is likely to result in missing data. In almost all trials with HRQOL endpoints, questionnaire completion rates, patient motivation, and patient numbers decrease over time (72),
- Moreover, some items, for example, those related to sexuality, generally have low response rates (in Europe). If such items are part of a total score, use of a separate sex questionnaire could be recommended in order not to reduce the power of the total score (94),
- In advanced-disease trials, some patients will be too ill to complete the questionnaire. Use of proxy or interviewer measures in this setting should be allowed, as these are preferable to no information (82), and
- The baseline HRQOL assessment can be part of the eligibility criteria. A minimal level of HRQOL impairment can be required so as to include patients with more severe disease and/or more motivated patients.

Description of Missing Data. The clinical report and any publication should provide a clear indication of the number of missing questionnaires and of missing items and their reasons (eg, noncompletion of a questionnaire due to death) in each group of patients (4,72,95). Several types of missing data are considered:

- **Missing at random (MAR):** missing items are equally distributed among all the items of the questionnaire, or missing questionnaires are equally distributed among the treatment groups,
- **Missing but not at random (MNAR):** the frequency of missing answers is particularly high for one or more items of the questionnaire (inferring a problem with their validity or acceptability) or the frequency of missing questionnaires is different between treatment groups or according to some other characteristic(s) of the patients, and
- It is difficult to confirm that missing data are really due to chance omission, independent from the treatment and/or the disease or other patient attributes; so considering missing data as **missing completely at random (MCAR)** is even more unlikely. In summary, there is strong evidence that missing data are not missing at random and cannot be ignored without introducing bias (72).

Imputation of Missing Data.

- Although several approaches to the analysis of missing data have been described, none of them has achieved consensus as a totally satisfactory method for the replacement of missing data (72,96). These techniques are not specific to HRQOL data. Whatever the method, the more missing data, the more the internal validity of the study will be questioned, especially if the differences in HRQOL scores between groups appear rather small, which is frequently the case. This is important even for high-quality data sets, given that 10% to 20% or less of missing single (index) or profile scores (due to missing questionnaires or missing items in one or several dimensions) may still cause sufficient bias in treatment comparisons to create problems when interpreting results (36,72),
- For missing items, a simple mean imputation is often used, based on the patient's own data (the missing item is replaced by the mean of all nonmissing items in the relevant scale), or on other patients' data (the missing item is replaced by the mean of that item for all patients who have responded) (97). Other techniques, such as regression imputation, are sometimes applied. Regression imputation involves regressing the missing item on the nonmissing items using data from subjects with

complete data; then the value of the missing item is predicted using the information from the completed items in the subject with the missing item (97). To perform imputation of missing items, at least 50% to 60% of items need to have been completed within a domain (70,91),
- For missing questionnaires or missing domains, LOCF is one of the techniques currently used. However, in some conditions, such as progressive diseases (eg, cancer), where missing data are unavoidable, more sophisticated techniques are recommended, including regression imputation, summary outcomes, growth-curve analysis, and mixed-effects models,
- Several approaches may be justified (sensitivity analysis), but if the results are not in agreement, then there is reason for concern, and
- In any case, authors should be explicit about what they have done, and should devote attention to the implication of their approach.

In summary, in order to reduce the quantity of missing HRQOL data, the decision to include an HRQOL questionnaire should be part of the initial design process and should be planned in the protocol. Administration, collection, and quality control should be subjected to the same rigorous standard procedures as for any traditional endpoint.

Multiple Statistical Tests

Statistical analysis of HRQOL data may be associated with a high incidence of type I error. Significant effects are found simply because of the multiple comparisons between treatment groups of many scales repeated over time (36). These comparisons within a study are not all independent. For example, performance of 10 statistical tests without adjustment on a questionnaire with 10 scales is associated with a type I error of 0.19. This means that a false statistically significant difference at the 0.05 level is likely to appear in approximately two scales at random.

There are several procedures that can be used to keep a global type I error at the level of 0.05 within a study. The procedure for controlling type I error should be prespecified in the protocol and documented in the report (83).

Reducing the Number of Statistical Comparisons.

- Some questionnaires allow the calculation of a single aggregated score across domains; thus, a single comparison test between groups is needed at the 5% significance level. However, a single index may reduce the information content for HRQOL and may mask or overestimate HRQOL treatment differences in important domains (6),
- Another method is to select those HRQOL endpoints that are expected to change (97). In practical terms this means selecting a few key domains in a questionnaire, or one questionnaire from a battery (7,36), and providing only descriptive statistics for the remaining domains. This procedure has several drawbacks. This selection has to be specified in advance in the protocol, and assessing HRQOL using a reduced number of domains will not allow investigators to draw conclusions about an HRQOL benefit,
- In the case of multiple assessments over time, a primary timepoint should be defined, and
- Another method for use with multiple assessments over time is to generate summary data, such as the average rate of change over time or area under the curve (AUC).

Statistical Adaptation to the Number of Comparisons. In order to compare all of the domains of a questionnaire (profile) without increasing the overall level of type I error, usually set at 0.05, several methods can be used, for example:

- The statistical tests which incorporate data from all timepoints (eg, repeated-measures ANOVA) can substitute for multiple comparisons at different timepoints (47),

- A lower level of significance can be set, adjusted to the number of statistical comparisons (eg, through use of the Bonferroni correction, where $\alpha = 0.05/k$, k being the number of comparisons) (97). However, this results in a focus on the smallest p-value and may yield conclusions that are counterintuitive. For example, an analysis of four endpoints (either four different timepoints or four HRQOL dimensions) may show effects in the same direction with a significance level of $p = 0.02$ for each endpoint. Using the Bonferroni adjustment, the differences would not be considered statistically significant because all values are greater than 0.0125 (0.05/4). In contrast, a statistically significant difference would be indicated if only one endpoint among the four reached a significance level of $p = 0.01$ (97). In a clinical trial, if all the comparisons show effects in the same direction and the internal validity of the trial is established, the conclusion should be favorable even if the p-values are over the adjusted level required. Thus, in some situations, when many or all of the domain changes are in the same direction, and provided that the design and analysis of the trial is appropriate, a less severe decision analysis tool (eg, Hochberg procedure) may be adequate even if the p-values of the changes are higher than the adjusted p-value for multiple comparisons (but lower than 0.05),
- A multivariate analysis including all dimensions of the questionnaire can be performed. Only if this analysis shows a statistically significant result is a univariate analysis performed on each dimension scale, and
- A growth-curve analysis can be performed.

Some authors advocate setting a less stringent level of significance (<0.05) in certain circumstances, for example, 0.10 (6). These infrequent cases should be very clearly defined and explained, and results should be taken with caution, because of the above mentioned risk of type I error, and of the doubtful relevance of a small difference that has reached the statistically significant level (98).

Recommendations

All of the following statistical issues should be prespecified and justified in the protocol before the start of the study, in order to minimize sources of bias and to increase confidence in the results and conclusions from the trial (83).

- *Study objectives,*
- *Hypotheses of superiority or equivalence,*
- *Use of HRQOL as primary or secondary endpoint,*
- *In the case of a profile instrument or a battery of instruments, selection of the major dimension(s) or instrument(s) which is/are expected to change,*
- *Sample size calculations:*
  - *When HRQOL is assessed in a clinical trial as a primary endpoint, a sample size must be calculated,*
  - *The same applies to HRQOL when used as a secondary endpoint: an a priori estimatation of expected changes in HRQOL scores must be calculated,*
  - *Without justification, too small a sample size is useless, and a large sample may demonstrate the statistical significance of a nonmeaningful difference, and*
  - *The rationale for the sample size determination needs to be documented in the protocol, the study report, and any publication,*
- *Scoring of the HRQOL questionnaire,*
- *Statistical tests planned, with the number of HRQOL comparisons (single score, profile with the number of domains, number of assessments over time),*
- *Methods for controlling for type I error,*
- *Missing data:*
  - *Procedures for prevention of missing data,*
  - *Description of missing data: Number of and reasons for missing items and missing questionnaires, and interpretation of whether missing data are random or not, and*
- *Methods for handling of missing data. In*

the case of a high number of missing data, a conservative procedure attitude (simple mean imputation) should be preferred.

INTERPRETATION OF RESULTS

- After a clinical trial has shown a statistically significant difference between two treatment groups on an HRQOL endpoint, the major issue is whether this difference is clinically relevant (99,100),
- Studies with common physiological measures allow clinicians to interpret what constitutes a meaningful clinical difference and to relate the results to standard norms (63, 101). Due to the lack of familiarity of most clinicians and policy-makers with HRQOL scores, there is no similar and intuitive interpretation of HRQOL scores. However, the availability of norm scores in different illnesses is growing, especially with some generic instruments such as the SF-36. This will help the interpretation of the results of clinical trials using the same instruments,
- Small differences in mean HRQOL scores between groups may be statistically significant, especially with large-scale clinical trials, but the clinical relevance may be difficult to interpret (2,100). For example, three questionnaires were used to evaluate the effect of one year of treatment with fluticasone propionate on functional status and sleep disturbance in children aged 4 to 11 years of age with asthma, and on the HRQOL of their parents. The mean score improved over the 52 weeks from 92.3 to 95.3 ($p < 0.05$) for functional status and from 89.2 to 90.6 ($p < 0.05$) for sleep disturbance, using a scale that ranged from 0 to 100 (102),
- It has been argued that any change in an HRQOL measure should be considered clinically significant, as any change represents a patient's perception of a modified health outcome. Several authors provide a definition of what they have termed 'minimal clinically important difference.' This has been defined as the smallest difference for a score in a domain of interest that patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management (99, 101). However, this does not directly suggest an operational method for defining clinical meaningfulness (88).

Several ways of assigning a qualitative meaning to an instrument's quantitative score have been proposed (2,3).

Some interpretations of HRQOL changes can be classified as **distribution based;** these are statistically derived (3,88,100):

- The magnitude of a between-group difference can be expressed as the **effect size,** which is the difference in score between groups divided by the pooled standard deviation of baseline scores. The effect size is called the standardized response mean if the denominator is the standard deviation of the difference in score. It is called the responsiveness statistic if the denominator is the standard deviation of the scores in stable patients (36,88,103), and
- Effect size may be a useful parameter for comparing scores in different studies or deciding on the relative importance of a treatment effect within a study. The generally accepted benchmarks are 0.20 for a small effect size, 0.50 for a moderate effect size, and 0.80 for a large effect size (103) (see Table 2),
- Effect size allows comparison of different questionnaires, which may help in interpretation, and
- Effect size is subject to the same criticism as statistical significance as it is dependent on the sample size and it does not provide an estimation of clinical significance (100).

Recently, it has been reported that the standard error of measurement could define clinical meaningfulness, as a change of one-standard error of measurement corresponded well to the patient-driven perception of change (104). This has to be confirmed by further studies.

Interpretation of HRQOL results can also be classified as **anchor based (or content**

TABLE 2
Effect Size and Relation to Change

Effect Size	No Change	Small Change (nonpertinent)	Moderate Change	Large Change
	<0.20	0.20≤–<0.50	0.50≤–<0.80	≥0.80

based). These interpretations are derived by comparison with other clinical changes or results (eg, global ratings, life events, threshold effect, changes with time, changes with therapy, disease conditions) (2,88,100). Such comparisons try to quantify the minimal important difference.

Examples of comparisons with global ratings:

- One study has used an anchor-based approach with subjects being asked to provide a rating of perceived change over time. The mean change of the EORTC Quality of Life Core Questionnaire (QLQ-C30) score (0–100 scale) was in the range 5 to 10 for those who reported 'little change,' in the range 10 to 20 for those who reported 'moderate change,' and >20 for those who reported 'very much change' (100). However, this approach has been criticized because of the retrospective assessment of change,
- Similarly, it has been suggested that a 5 to 10 point change (on a 0–100 scale) along any of the eight scales of the SF-36 is clinically meaningful (105),
- Pooled results during validation studies using the Chronic Respiratory Questionnaire and the Chronic Heart Failure Questionnaire suggested that a change in mean score of 0.5 on a seven-point Likert scale may be considered the smallest difference that the patient perceives as beneficial and hence can be considered the minimal important difference when compared to a 15-point global rating scale of change (ranging from −7 to +7). A within-subject change in HRQOL score of 0.5 represents the minimal important difference, a change in score of 1.0 may be considered a moderate change, and a change in score >1.5 is likely to represent a large change (29,98–101) (see Table 3). Thus, in a domain with four items, patients will consider a one-point change in two or more items important (98); or for a six-item domain, a change of three to four points in the domain score is considered small, a five- to six-point change is moderate and a change of ≥7 points is large. This approach has also been criticized as the comparison is made with a global rating that is not validated (ie, how can a change from 1 to 3 on a global rating be defined as a small change?),
- Another example of estimating the minimal important difference through an anchor-based interpretation is that a change of 0.02 points on the 'Quality of Well-Being Scale' is equivalent in all treated patients with rheumatoid arthritis improving from moving their own wheelchair without help to walking with physical limitation (106),
- In a recent case, results of a randomized clinical trial were anchored to the difference observed over time in a cohort of similar patients in an observational study. The difference seen in the clinical trial was roughly comparable to several months of natural progression of the disease (data not published), and
- Another interpretation is to translate an HRQOL improvement into a clinical change. For example, the top score of the SF-36 physical functioning scale is achieved only by being able to perform all physical activities measured. An improvement in this scale (range: 0–100) from 40 to 50 corresponds to 18% more people indicating that they can walk one block without limitation (107). One limitation is that the interpretation of a 10-point improvement may not be the same at another level of the scale.

Estimating the clinically important difference seems critical for judging the magnitude of the benefit when comparing two treatments, calculating sample size, making inferences about the percentage of patients improved by a therapeutic intervention (98,103) or calculating the number of patients needed to treat (98).

The **number of patients needed to treat (NNT)** is another useful way of presenting results that is becoming more frequently used with clinical outcomes. The NNT is obtained from the reciprocal of the absolute risk reduction (108,109). In practical terms, NNT is the number of patients that have to be treated during a defined period for one patient to improve his HRQOL (responder). This number is directly interpretable by clinicians. For example, consider a situation in which a randomized clinical trial has shown a mean difference of 0.25 between HRQOL scores and in which the clinically important difference is set at 0.50. Of the patients treated with the new drug, 25% improved by a magnitude of ≥0.50, while no patients in the control group showed any improvement. The number of patients who obtain important benefit from treatment is the number of patients with a difference of ≥0.50 favoring the treatment period, minus the number of patients with a difference of ≥0.50 or more favoring the control period. The reciprocal of this difference is the NNT. In this case, the NNT for one patient to improve is four (110). The major limitation of this technique is defining the status of responder/nonresponder (ie, how to decide the clinically important difference?).

Some interpretation still remains unclear:

- Does a change in an HRQOL score have the same value at the top and the bottom of the scale?
- If a statistically significant and clinically meaningful difference exists between two groups on only two of eight or nine subscales, would it be possible to claim that the drug improves HRQOL (9)? What is the threshold of the number of scales in a questionnaire for which a significant difference should be observed in order to claim a therapeutic improvement of HRQOL? At least two of the three major scales reflecting physical, psychological, and social functions should show a significant difference in score, unless, depending on the disease or condition, one or more important scales fail(s) to show a difference between treatment groups or show(s) a deterioration with the drug under evaluation. For exam-

TABLE 3
Estimation of a Minimal Important Difference According to the Global Rating Change

Answer to the Global Rating Change[1] (98)	Worse	Better	Interpretation of Change	Mean Change in HRQOL Scale
A very great deal	–7	+7	Large	1.5
A great deal	–6	+6		
A good deal	–5	+5	Moderate	1
Moderately	–4	+4		
Somewhat	–3	+3	Small	0.5
A little	–2	+2		
Almost the same	–1	+1		
About the same	0		None	

[1]"Overall, has there been any change in your shortness of breath during your daily activities since the last time you saw us?"

ple, a drug evaluated in arteriopathy improved some of the scales of a specific questionnaire, but not the pain scale, which is one of the most important in this condition (111),
- In a questionnaire with several scales, where a few have been specified and justified a priori as the key endpoints, if significant differences are shown for these scales, a conclusion of HRQOL improvement is acceptable, providing that at least two of the three major scales reflecting physical, psychological, or mental and social functions are among these scales,
- A statistically significant change in only one or two HRQOL domain(s) among several does not reflect HRQOL. At most, it reflects the domain(s) concerned, but it could be argued that the result is due to type I error. The claim should be restricted to the domains enhanced and not to HRQOL improvement, and
- What can be concluded when some domain scores improve in a treatment group versus the comparator while other domain scores deteriorate or while symptom scores do not change or worsen? It is reasonable to assume that HRQOL results should not be in conflict with clinical results.

Negative Results

There are several possible explanations for a nonsignificant difference between two groups in a clinical trial:

- Patients' disease is not severe enough (eligibility criteria are incorrect),
- The HRQOL questionnaire is not sufficiently reliable, valid, and/or responsive,
- The sample size is too small, or
- The treatment under study does not improve HRQOL,

The absence of difference between two treatment groups does not always mean that there is no difference between the treatments (112). An underpowered study (ie, too few patients included) or a study using an unresponsive questionnaire will lead to a no significant difference, but it would be incorrect to claim equivalence in such a study (7). In an equivalence trial, evidence of responsiveness of the selected instrument must be acquired before the trial.

Recommendations

In the absence of definitive guidelines on the interpretation of HRQOL scores, some (at least the first four points, but ideally all, especially if HRQOL is the primary endpoint) of the following data should be presented:

- *Clear description of the content of domains,*
- *Clear distribution of HRQOL scores within and between groups,*
- *Ninety five percent confidence interval of the difference and/or odds ratio of the difference,*
- *Effect size and/or standardized response mean,*
- *Relationship of scores to clinical endpoints (if measured during the trial),*
- *Comparisons of scores with norms derived from different disease groups and the general population (if available) and/or scores obtained in other studies in a similar population, to estimate a clinical significance,*
- *Comparison with external criteria (clinical endpoints, global ratings, resource utilization, etc.) to estimate a clinical significance, and*
- *Number needed to treat.*

CONCLUSIONS

Even if some major issues with regard to HRQOL are still unresolved (eg, missing data, interpretation of HRQOL results through the minimal important difference), regulatory authorities will more readily accept HRQOL results that are statistically significant if they have confidence in the quality of the trial itself and find the information and justifications described above in the protocol and clinical report.

Many drugs are approved based upon a statistically significant difference between

groups of a physiological or clinical endpoint, the clinical relevance of which remains unknown. The problem for HRQOL results is that, as they are more difficult to interpret, reviewers of the dossier need to have more trust in the internal validity of the trial than for the other endpoints.

A carefully designed and implemented randomized controlled clinical trial, with HRQOL as a primary endpoint, a sample size calculation, data quality control procedures, and rigorous statistical analysis, which yielded statistically significant differences in favor of the treatment under study, should be completely acceptable (3). The question is, how many such HRQOL studies are required for labeling or for a promotional/advertising claim? Would a single clinical trial, incorporating rigorous and valid HRQOL assessment as the primary endpoint, be enough to support a claim (6)? Or would two clinical trials be necessary when HRQOL parameters are considered secondary endpoints (7)? Whatever the situation, it is likely that evidence of significant and concordant results in more traditionally accepted endpoints, such as symptom severity, will still be required until more experience is gained in the significance of HRQOL.

Finally, one way to improve HRQOL assessment in clinical trials is to plan it earlier in the clinical development program of a new drug. Implementing HRQOL assessment in Phase II clinical trials allows verification of the psychometric properties of an HRQOL questionnaire, provides preliminary data, and allows more precise hypotheses to be tested during Phase III clinical trials. Whether HRQOL is considered a primary or secondary endpoint, the scientific principles of clinical trial design must apply to HRQOL.

APPENDIX: CHECKLIST FOR DESIGNING, CONDUCTING, AND REPORTING HRQOL STUDIES IN CLINICAL TRIALS

This checklist is intended for use principally when HRQOL is a primary endpoint, but the same standards should apply when HRQOL is a secondary endpoint. For a description of each of the items, please refer to the indicated sections within this document.

Decision Making Rules

In the checklist, it is mandatory (***'M'***) to answer 'Yes' for some of the items. For the rest of the items, most answers should be 'Yes' for there to be confidence in the results of the trial. The items marked by an ***'I'*** are of greatest importance.

	Yes/No	Comment
1. **Is the study design clearly described? (methodological design)** It is assumed that the methodological principles of randomized clinical trials are fulfilled and clearly reported: that is, type of study design, blindness rules, method of randomization, choice of comparator, statistical analysis plan, description of the follow-up of patients.		
1.1. Is it a comparative study?	***M***	
1.2. Is the method of randomization adapted and described?	***M***	
1.3. Are double- or single-blind design or procedures adopted to ensure a minimum of bias when administering the HRQOL questionnaire?	***M***	
2. **Is the scope and definition of the HRQOL evaluation adequately justified?** *(Sections: Added value of patient-based HRQOL assessment; Selection of an HRQOL instrument)*	***I***	
2.1. Is the relevance of assessing HRQOL justified for this particular disease? Are specifics of the trial design and study population taken into account?	***I***	
2.2. Is the choice of the HRQOL questionnaire(s) documented?	***I***	
2.3. Does the questionnaire present evidence of statistical validation?	***I***	
2.4. Does the questionnaire cover the domains that are expected to change?	***I***	

	Yes/No	Comment
2.5. Are the research objectives of the HRQOL component clearly stated?	***I***	
2.6. Are specific, testable HRQOL hypotheses posed (eg, which are the domains of interest? Which size effect or clinically important difference is expected?)	***I***	
2.7. Is HRQOL defined as a primary or secondary endpoint in the trial?	***I***	
3. **Is there a clear description of, and rationale for, the following study design elements as they relate to the HRQOL component of the trial?**		
(Section: Study design—practical considerations)		
3.1. Sampling of patients: description of the method used. Is the sample representative of (1) the study population; (2) the patients who are likely to receive the treatment?		
3.2. Sampling of centers (in the case of multicenter trials).		
3.3. Eligibility criteria.		
3.4. Timing and frequency of HRQOL assessment.		
3.5. Mode and site of HRQOL administration.		
3.6. Data monitoring and quality assurance.	***I***	
3.7. Procedures for prevention and handling of missing data.	***I***	
Are the properties of the HRQOL measure(s) adequately described and documented?		
(Section: Instrument properties)		
3.8. Number of items and domains.	***I***	
3.9. Instrument scaling and scoring.	***I***	
3.10. Reliability (internal consistency, test–retest).	***I***	
3.11. Validity (content, structural, clinical, concurrent, predictive).	***I***	
3.12. Responsiveness.		
3.13. Respondent burden: time needed.		
3.14. Cultural adaptation.	***I***	
3.15. Evidence of psychometrics for translated questionnaires.	***I***	
3.16. If the psychometric validation is not completed at time of study implementation, is a validation plan provided?	***I***	
4. **Is a detailed statistical analysis plan of the HRQOL data provided?**		
(Section: Statistical issues)		
4.1. Superiority or equivalence trial.	***M***	
4.2. Sample size and statistical power.	***M***	
4.3. Intent-to-treat analysis (ITT) or modified ITT (provide criteria).	***M***	
4.4. Descriptive and inferential statistics for evaluating changes over time and/or group differences.	***I***	
4.5. Procedures to maintain or preserve the overall Type I error rate at a specific level (eg, 0.05) in the case of multiple statistical tests.	***I***	
4.6. Imputation of missing data (items and questionnaires).	***I***	
5. **Reporting and interpretation of results**		
(Sections: Statistical issues; Interpretation of results)		
5.1. Is the following information provided?		
5.1.1. Participation rate (at study entry and during follow-up).	***I***	
5.1.2. Demographic and medical characteristics of the HRQOL study population.	***I***	
5.1.3. Data completeness (ie, missing questionnaires and missing items).	***I***	
5.2. Are the results presented in accordance with the original statistical analysis plan?	***I***	
5.2.1. If not, are deviations explained and justified?	***I***	
5.3. Is there an attempt to interpret the statistical results in terms of clinical significance?		
5.3.1. Distribution of HRQOL scores within and between groups.		
5.3.2. 95% confidence interval of the difference and/or odds ratio.		
5.3.3. Effect size and/or standardized response mean.		
5.3.4. Comparisons of scores with norm (if available) and/or scores obtained in other studies in a similar population to enable estimation of clinical significance.		
5.3.5. Comparison with external criteria (clinical endpoints, global rating, etc.) to estimate a clinical significance.		
5.3.6. Number needed to treat.		

REFERENCES

1. Johnson JR, Temple R. Food and Drug Administration requirements for approval of new anticancer drugs. *Cancer Treat Rep.* 1985;69:1155–1159.
2. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, Roberts JS. Evaluating quality-of-life and health status instruments: Development of scientific review criteria. *Clin Ther.* 1996;18:979–992.
3. Scientific Advisory Committee. Instrument Review Criteria. *Med Outcomes Trust Bull*. 1995;3:I–IV.
4. Staquet M, Berzon R, Osoba D, Machin D. Guidelines for reporting results of quality of life assessments in clinical trials. *Qual Life Res*. 1996;5:496–500.
5. Fayers PM, Hopwood P, Harvey A, Girling DJ, Machin D, Stephens R, on Behalf of the MRS Cancer Trials Office. Quality of Life Assessment in Clinical Trials—Guidelines and a Checklist for Protocol Writers: The UK Medical Research Council Experience. *Eur J Cancer.* 1997;33:20–28.
6. Revicki DA, Leidy NK, Geneste B. Recommendations for evaluating the validity of quality of life claims for labeling and promotion. *Value Health*. 1999;2:113–127.
7. Smith N. Quality of life studies from the perspective of an FDA reviewing statistician. *Drug Inf J*. 1993;27:617–623.
8. Sanders C, Egger M, Donovan J, Tallon D, Frankel S. Reporting on quality of life in randomised controlled trials: Bibliographic study. *Br Med J*. 1998; 317:1191–1194.
9. Chassany O, Bergmann JF, Caulin C. Reporting on quality of life in randomised controlled trials. *Br Med J*. 1999;318:1142.
10. Kong SX, Gandhi SK. Methodologic assessments of quality of life measures in clinical trials. *Ann Pharmacother.* 1997;31:830–836.
11. Brown RS. Strategies and pitfalls in quality of life research. *Hepatology.* 1999;29(Suppl. 1):9S–12S.
12. de Haes J, Curran D, Young T, Bottomley A, Flechtner H, Aaronson N, Blazeby J, Bjordal K, Brandberg Y, Greimel E, Maher J, Sprangers M, Cull A. Quality of life evaluation in oncological clinical trials—the EORTC model. The EORTC quality of life study group. *Eur J Cancer.* 2000;36:821–825.
13. Chassany O. Does dehydroepiandrosterone improve well-being? *Presse Med.* 2000;29:1354–1355.
14. Schipper H, Clinch J, Powell V. Definition and Conceptual Issues. In: Spilker B, ed. *Quality of Life Assessments in Clinical Trials*. New York, NY: Raven Press; 1990:11–24.
15. Spitzer WO. State of science 1986: Quality of life and functional status as target variables for research. *J Chronic Diseases*. 1987;6:465–471.
16. Sullivan M, Karlsson J, Taft C. How to Assess Quality of Life in Medicine: Rationale and Methods. In: B Guy-Grand, G Ailhaud, eds. *Progress in Obesity Research:8.* London: Libbey; 1999:749–755.
17. World Health Organization. *World Health Organization Constitution. Basic Documents*. Geneva, Switzerland: World Health Organization; 1948.
18. Patrick DL, Erickson P. *Health Status and Health Policy: Allocating Resources to Health Care*. New York, NY: Oxford University Press; 1993.
19. Burke L. Acceptable evidence for pharmaceutical advertising and labeling, DIA Workshop on Pharmacoeconomic and Quality of Life Labeling and Marketing Claims, October 3, 2000, New Orleans, LA.
20. Aaronson NK. Quality of life assessment in clinical trials: Methodologic issues. *Control Clin Trials*. 1989;10:S195-S208.
21. Testa MA, Simonson DC. Assessment of quality of life outcomes. *N Engl J Med*. 1996;334:835–840.
22. Revicki DA, Ehreth JL. Health-related quality-of-life assessment and planning for the pharmaceutical industry. *Clin Ther.* 1997;19:1101–1115.
23. Feinstein AR. Clinimetric perspectives. *J Chronic Diseases*. 1987;40:635–640.
24. Bulpitt CJ, Fletcher AE. The measurement of quality of life in hypertensive patients: A practical approach. *Br J Clin Pharmacol*. 1990;30:353–364.
25. Wiklund I, Dimenäs E, Wahl M. Factors of importance when evaluating quality of life in clinical trials. *Control Clin Trials*. 1990;11:169–179.
26. Wiklund I. Aspects of quality of life in gastrointestinal disease: Some methodological issues. *Scand J Gastroenterol*. 1995;208:129–132.
27. Trus TL, Laycock WS, Branum GD, Waring JP, Hunter JG. Quality of life scores correlate poorly with subjective and objective measurements of gastroesophageal reflux. *Gastroenterol.* 1997;112:A1480.
28. Guyatt GH, Thompson PJ, Berman LB, Sullivan MJ, Townsend M, Jones NL, Pugsley SO. How should we measure function in patients with chronic heart and lung disease? *J Chronic Diseases*. 1985;38:517–524.
29. Rector TS, Kubo SH, Cohn JN. Patients' Self-assessment of their cognitive heart failure: II: Content, reliability and validity of a new measure. The Minnesota Living with Heart Failure Questionnaire. *Heart Failure*. 1987;3:198–209.
30. Juniper EF, Guyatt GH, Ferrie PJ, Griffith LE. Measuring quality of life in asthma. *Am Rev Respir Dis*. 1993;147:832–838.
31. Wiklund I, Comerford MB, Dimenas E. The relationship between exercise tolerance and quality of life in angina pectoris. *Clin Cardiol*. 1991;14:204–208.
32. Guyatt GH, Feeny DH, Patrick DL. Measuring Health-related quality of life. *Ann Int Med*. 1993; 118:622–629.
33. Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, McGlynn EA, Ware JE. Functional status and well-being of patients with chronic conditions. Results from the Medical Outcomes Study. *JAMA*. 1989;262:907–913.
34. Bayliss MS, Gandek B, Bungay KM, Sugano D,

Hsu MA, Ware JE. A questionnaire to assess the generic and disease-specific health outcomes of patients with chronic hepatitis C. *Qual Life Res*. 1998; 7:39–55.
35. Ganz PA, Coscarelli A, Fred C, Kahn B, Plinski ML, Petersen L. Breast cancer survivors: Psychological concerns and quality of life. *Breast Cancer Res Treat*. 1996;38:183–199.
36. Fletcher A, Gore S, Jones D, Fiztpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care. II: Design, analysis, and interpretation. *Br Med J*. 1992;305:1145–1148.
37. Gill TM, Feinstein AR. A critical appraisal of the quality of quality of life measurements. *JAMA*. 1994;272:619–626.
38. Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. *Med Care*. 1989;27(Suppl 3):S217–S232.
39. Goldstein RS, Gort EH, Guyatt GH, Stubbing D, Avendano MA. Prospective randomized controlled trial of respiratory rehabilitation. *Lancet*. 1994;344: 1394–1397.
40. Tandon PK, Stander H, Schwarz RPJ. Analysis of quality of life data from a randomized, placebo controlled heart-failure trial. *J Clin Epidemiol*. 1989;42:955–962.
41. Laupacis A, Wong C, Churchill D. The use of generic and specific quality-of-life measures in hemodialysis patients treated with Erythropoietin. *Control Clin Trials*. 1991;121:68S–79S.
42. Galmiche JP, Barthelemy P, Hamelin B. Treating the symptoms of gastro-oesophageal reflux disease: A double-blind comparison of Omeprazole and Cisapride. *Aliment Pharmacol Ther*. 1997;11: 765–773.
43. Tarter RE, Switala J, Arria A, Plail J, Van Thiel D. Quality of life before and after orthotopic hepatic transplantation. *Arch Int Med*. 1991;151:1521–1526.
44. Chassany O, Marquis P, Scherrer B, Read NW, Finger T, Bergmann JF, Fraitag B, Geneve J, Caulin C. Validation of a specific quality of life questionnaire in Functional Digestive Disorders (FDDQL). *Gut*. 1999;44:527–533.
45. Bergner M, Rothman ML. Health status measures: an overview and guide for selection. *Ann Rev Public Health*. 1987;8:191–210.
46. Wiklund I. Quality of life and regulatory issues. *Scand J Gastroenterol*. 1996;31(Suppl. 221):37–38.
47. Testa MA, Nackley JF. Methods for quality-of-life studies. *Ann Rev Public Health*. 1994;15:535–559.
48. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: What are the necessary measurement properties? *J Clin Epidemiol*. 1992;45:1341–1345.
49. Hays R, Anderson R, Revicki DA. Assessing Reliability and Validity of Measurement in Clinical Trials. In: Staquet MJ, Hays RD, Fayers PM, eds. *Quality of Life Assessment in Clinical Trials*. New York, NY: Oxford University Press; 1998: 169–182.
50. Fletcher A. Quality-of-life measurements in the evaluation of treatment: Proposed guidelines. *Br J Clin Pharmacother.* 1995;39:217–222.
51. Marquis P, Fayol C, Joire JE, Leplège A. Psychometric properties of a specific quality of life questionnaire in angina pectoris patients. *Qual Life Res*. 1995;4:540–546.
52. Jenkinson C, Ziebland S, Fitzpatrick R, Mowat A. Sensitivity to change of weighted and unweighted versions of two health status measures. *Intl J Health Sciences*. 1991;2:189–194.
53. Zimmerman M. Weighted versus unweighted life event scores: Is there a difference? *J Human Stress*. 1983;9:30–35.
54. Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measures in health care. I: Applications and issues in assessment. *Br Med J*. 1992;305:1074–1077.
55. Kirschner B, Guyatt GH. A methodologic framework for assessing health indices. *J Chronic Diseases*. 1985;38:27–36.
56. Guyatt G, Mitchell A, Irvine EJ, Singer J, Williams N, Goodacre R, Tompkins C. A new measure of health status for clinical trials in inflammatory bowel disease. *Gastroenterol*. 1989;96:804–810.
57. Eypasch E, Williams JI, Wood-Dauphinee S, Ure BM, Schmülling C, Neugebauer E, Troidl H. Gastrointestinal Quality of Life Index : Development, validation and application of a new instrument. *Br J Surg*. 1995;82:216–222.
58. Scharloo M, Kaptein AA, Weinman JA, Willems LNA, Rooijmans HGM. Physical and psychological correlates of functioning in patients with chronic obstructive pulmonary disease. *J Asthma*. 2000;37:17–29.
59. Faller H, Bulzebruck H, Drings P, Lang H. Coping, distress, and survival among patients with lung cancer. *Arch Gen Psychiatry.* 1999;56:756–762.
60. Stockler MR, Osoba D, Corey P, Goodwin PJ, Tannock IF. Convergent, discriminative, and predictive validity of the Prostate Cancer Specific Quality of Life Instrument (PROSQOLI) assessment and comparison with Analogous Scales Form the EORTC QLQ-C30 and a trial-specific module. *J Clin Epidemiol*. 1999;52:653–666.
61. Feagan BG, McDonald JW, Rochon J, Laupacis A, Fedorak RN, Kinnear D, Saibil F, Groll A, Archambault A, Gillies R, Valberg B, Irvine JE, for the Canadian Crohn's Relapse Prevention Trial Investigators. Low-dose Cyclosporine for the treatment of Crohn's Disease. *N Engl J Med*. 1994;330:1846–1851.
62. Guyatt GH, Walter S, Norman G. Measuring health status over time. Assessing the usefulness of evaluative instruments. *J Chronic Diseases*. 1987;40: 171–178.
63. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care*. 1989;27(Suppl. 3):S178–S189.
64. Deyo RA, Diehr P, Patrick DL. Reproducibility and

responsiveness of health status measures. Statistics and strategies for evaluation. *Control Clin Trials*. 1991;12:S142–S158.
65. Acquadro C, Jambon B, Ellis D, Marquis P. Languages and Translation Issues. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials*. Second edition. Philadelphia, PA: Lippincott-Raven Publishers; 1996:575–585.
66. Bullinger M, Anderson R, Cella D, Aaronson N. Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. *Qual Life Res*. 1993;2:451–459.
67. Hunt SM. Cross-cultural Issues in the Use of Quality of Life Measures in Randomized Clinical Trials. In: Staquet MJ, Hays RD, Fayers PM, eds. *Quality of Life Assessment in Clinical Trials: Methods and Practice*. New York, NY: Oxford University Press; 1998: 51–67.
68. Bullinger M, Alonso J, Apolone G, Leplege A, Sullivan M, Wood-Dauphinee S, Gandek B, Wagner A, Aaronson N, Bech P, Fukuhara S, Kaasa S, Ware JE Jr. Translating health status questionnaires and evaluating their quality: the IQOLA project approach. International Quality of Life Assessment. *J Clin Epidemiol*. 1998;51:913–923.
69. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *J Clin Epidemiol*. 1993;46:1417–1432.
70. Gandek B, Ware JE Jr, Aaronson NK, Alonso J, Apolone G, Bjorner JB, Brazier J, Bullinger M, Fukuhara S, Kaasa S, Leplege A, Sullivan M. Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countries: Results from the IQOLA project. *J Clin Epidemiol*. 1998;51:1149–1158.
71. Spilker B. Quality of Life Trials. In: Spilker B, ed. *Guide to Clinical Trials*. New York, NY: Raven Press; 1991:370–378.
72. Bernhard J, Cella DF, Coates CA, Fallowfield L, Ganz PA, Moinpour CM, Mosconi P, Osoba D, Simes J, Hürny C. Missing quality data in cancer clinical trials: Serious problems and challenges. *Stat Med*. 1998;17:517–532.
73. Osoba D. Rationale for the timing of health-related quality-of-life assessments in oncological palliative therapy. *Cancer Treat Rev.* 1996;22(Suppl.A):69–73.
74. Rothman ML, Hedrick SC, Bulcroft KA, Hickam DH, Rubenstein LZ. The validity of proxy-generated scores as measures of patient health status. *Med Care*. 1991;29:115–124.
75. Sneeuw KCA, Aaronson NK, Osoba D, Muller MJ, Hsu MA, Yung WK, Brada M, Newlands ES. The use of significant others as proxy raters of the quality of life of patients with brain cancer. *Med Care*. 1997;35:490–506.
76. Sprangers MAG, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *J Clin Epidemiol*. 1992;45:743–760.
77. Testa MA, Hollenberg NK, Anderson RB, Williams GH. Assessment of quality of life by patient and spouse during antihypertensive therapy with Atenolol and Nifedipine gastrointestinal therapeutic system. *Am J Hypertens*. 1991;4:363–373.
78. Mathias SD, Bates MM, Pasta DJ, Cisternas MG, Feeny D, Patrick DL. Use of the health utilities index with stroke patients and their caregivers. *Stroke*. 1997;28:1888–1894.
79. Sneeuw KCA, Aaronson NK, Sprangers MAG, Detmar SB, Wever LDV, Schornagel JH. Value of caregiver ratings in evaluating the quality of life of patients with cancer. *J Clin Oncol*. 1997;15:1206–1217.
80. Sneeuw KCA, Aaronson NK, de Haan RJ, Limburg M. Assessing quality of life after stroke. The value and limitations of proxy ratings. *Stroke*. 1997;28:1541–1549.
81. Stephens RJ, Hopwood P, Girling DJ, Machin D. Randomized trials with quality of life endpoints. Are doctor's ratings of patients' physical symptoms interchangeable with patients' self-ratings? *Qual Life Res*. 1997;6:225–236.
82. Simes RJ, Greatorex V, Gebski VJ. Practical approaches to minimize problems with missing quality of life data. *Stat Med*. 1998;17:725–737.
83. Statistical Principles for Clinical Trials. ICH Harmonised Tripartite Guideline. CPMP/ICH/363/96. *http://www.eudra.org/humandocs/humans/ICH.htm*.
84. Lydick E, Epstein RS. Clinical Significance of Quality of Life Data. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials*. Philadelphia, PA: Lippincott-Raven Publishers; 1996: 461–465.
85. Chassany O, Duracinsky M. Ethics and clinical trials. *Fundam Clin Pharmacol*. 1999;13:437–444.
86. Rush DR, Stelmach J, Young TL, Kirchdoefer LJ, Scott-Lennox J, Holverson HE, Sabesin SM, Nicholas TA. Clinical effectiveness and quality of life with ranitidine vs placebo in gastroesophageal reflux disease patients: a clinical experience network (CEN) Study. *J Fam Pract*. 1995;41:126–136.
87. Juniper EF. Quality of life questionnaires: Does statistically significant = clinically important? *J Allergy Clin Immunol*. 1998;102:16–17.
88. Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual Life Res*. 1993;2:221–226.
89. Offerhaus L. Measurement of the quality of life in clinical trials: in pursuit of the unapproachable. *Eur J Clin Pharmacol*. 1991;40:205–208.
90. Curran D, Fayers PM, Molengerghs G, Machin D. Analysis of Incomplete Quality of Life Data in Clinical Trials. In: Staquet MJ, Hays RD, Fayers PM, eds. *Quality of Life Assessment in Clinical Trials*. New York, NY: Oxford University Press; 1998:249–280.
91. Fairclough D, Peterson H, Cella D, Bonomi P. Comparison of several model-based methods for analyzing incomplete quality of life data in cancer clinical trials. *Stat Med*. 1998;17:781–796.

92. Hayden KA, Moinpour CM, Metch B, Feigl P, O'Brian RM, Green S, Osborne CK. Pitfalls in quality of life assessment: Lessons from a Southwest Oncology Group breast cancer clinical trial. *Oncol Nurs Forum.* 1993;20:1415–1419.
93. Olschewski M. Compliance with QoL assessment in multicentre German breast cancer trials. *Stat Med*. 1998;17:571–575.
94. Wiklund I, Junghard O, Grace E, Talley NJ, Kamm M, Veldhuyzen van Zanten S, Paré P, Chiba N, Leddin DS, Bigard MA, Colin R, Schoenfeld P. Quality of life in reflux and dyspepsia patients. Psychometric documentation of a new disease-specific questionnaire (QOLRAD). *Eur J Surg*. 1998;83(Suppl.583):41–49.
95. Machin D, Weeden S. Suggestions for the presentation of quality of life data from clinical trials. *Stat Med*. 1998;17:711–724.
96. Troxel AB, Fairclough DL, Curran D, Hahn EA. Statistical analysis of quality of life with missing data in cancer clinical trials. *Stat Med.* 1998;17: 517–532.
97. Fairclough DL, Gelber RD. Quality of Life: Statistical Issues and Analysis. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials*. Philadelphia, PA: Lippincott-Raven Publishers; 1996:427–435.
98. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *Br Med J*. 1998;316:690–693.
99. Jaeschke R, Singer J, Guyatt GH. Measurement of Health Status: Ascertaining the Minimal Clinically Important Difference. *Control Clin Trials*. 1989; 10:407–415.
100. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol*. 1998; 16:139–144.
101. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol*. 1994;47:81–87.
102. Mahajan P, Pearlman D, Okamoto L. The effect of Fluticasone Propionate on functional status and sleep in children with asthma and on the quality of life of their parents. *J Allergy Clin Immunol*. 1998;102:19–23.
103. Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J, Matchar D. Determining clinically important differences in health status measures. *PharmacoEconomics*. 1999;15:141–155.
104. Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Med Care*. 1999; 37:469–478.
105. Ware JE, Snow KK, Kosinski M, Gandek B. *SF-36 Health Survey: Manual and Interpretation Guide*. Boston, MA: The Health Institute, New England Medical Center; 1993.
106. Thompson MS, Read JL, Hutchings HC, Paterson M, Harris ED. The cost effectiveness of Auranofin: Results of a randomized clinical trial. *J Rheumatol*. 1988;15:35–42.
107. Ware JE, Keller SD. Interpreting General Health Measures. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials*. Philadelphia, PA: Lippincott-Raven Publishers; 1996:445–460.
108. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *Br Med J*. 1995;310:452–454.
109. Altman DG. Confidence intervals for the number needed to treat. *Br Med J*. 1998;317:1309–1312.
110. Juniper EF. The value of quality of life in asthma. *Eur Respir Rev*. 1997;49:333–337.
111. Liard F, Benichou AC, Gamand S, Lehert P. The effects of Naftidrofuryl on quality of life. *Disease Manage Health Outcomes*. 1997;2(Suppl.1):71–78.
112. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *Br Med J*. 1995;311:485–486.