Measuring disease-specific quality of life in clinical trials

Gordon H. Guyatt,*† MD
Claire Bombardier,‡
Peter X. Tugwell,* MD

While measurement of quality of life is a vital part of assessing the effect of treatment in many clinical trials, a measure that is responsive to clinically important change is often unavailable. Investigators are therefore faced with the challenge of constructing an index for a specific condition or even for a single trial. There are several stages in the development and testing of a quality-of-life measure: selecting an initial item pool, choosing the "best" items from that pool, deciding on questionnaire format, pretesting the instrument, and demonstrating the responsiveness and validity of the instrument. At each stage the investigator must choose between a rigorous, time-consuming approach to questionnaire construction that will establish the clinical relevance, responsiveness and validity of the instrument and a more efficient, less costly strategy that leaves reproducibility, responsiveness and validity untested. This article describes these options and outlines a pragmatic approach that yields consistently satisfactory disease-specific measures of quality of life.

Si on reconnaît l'importance, dans beaucoup d'essais cliniques, de déterminer la qualité de la vie en rapport avec le traitement, on manque souvent d'un instrument de mesure capable de refléter des modifications importantes de l'état clinique. Dès lors le chercheur s'attache à établir un indice valable pour une maladie donnée, voire pour tel essai. La conception et la mise à l'épreuve d'un instrument de mesure de la qualité de la vie comportent plusieurs stades: réunir des critères, en choisir les meilleurs, concevoir un questionnaire, le soumettre à un essai préliminaire, et en déterminer la sensibilité et la validité. À chacun de ces stades le chercheur doit choisir entre une démarche rigoureuse et de longue haleine permettant de bien déterminer la pertinence clinique de l'instrument, sa sensibilité et sa validité d'une part, et de l'autre une démarche moins coûteuse qui ne permet pas de juger de ces qualités. On décrit ici les choix qui sont à faire et des moyens pratiques de créer pour une maladie donnée un instrument capable de mesurer de façon constamment fidèle la qualité de la vie.

From *the departments of Clinical Epidemiology and Biostatistics, and Medicine, McMaster University Health Sciences Centre, Hamilton, Ont., and ‡the Department of Medicine, University of Toronto

†Career scientist, Ontario Ministry of Health

Reprint requests to: Dr. Gordon H. Guyatt, Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre, Rm. 3H7, 1200 Main St. W, Hamilton, Ont. L8N 3Z5

Many medical interventions are designed to improve the quality rather than extend the duration of the patient's life. A direct measure of quality of life is required to assess the benefit of such interventions. While a number of quality-of-life instruments have been developed for the general population,[1-5] they are unlikely to detect small, clinically important changes. Therefore, investigators have developed disease-specific instruments for patients with cancer,[6-8] joint diseases,[9-12] heart disease[13] and chronic lung disease.[14] While these instruments, once validated, are likely to be useful to other investigators, they may have a narrow range of applicability. Different illnesses may affect different bodily functions and lead to different physical and emotional problems. For example, one recently developed instrument tests the relative toxicity of two regimens for managing stage II carcinoma of the breast,[15] but it is only applicable to chemotherapeutic regimens with similar toxicity. Also, some instruments may have a narrow range of applicability because different treatments are designed to ameliorate impairment of different aspects of quality of life. In planning a clinical trial of an occupational therapy program for patients with rheumatoid arthritis, for example, we found existing questionnaires for assessing the effects of anti-inflammatory therapy were inappropriate, so a special questionnaire was required.

These examples illustrate that a new instru-

ment will often be required to measure the effect of a specific treatment on quality of life. Clinical investigators who are inexperienced in questionnaire development have responded to this dilemma by creating ad-hoc measures.[16,17] The results of their studies are difficult to interpret because of failure to attend appropriately to what patients consider important and to issues such as clinical credibility, reproducibility, responsiveness and validity.

In this review we present a strategy for developing questionnaires to measure disease-specific quality of life. It is based on previous work that identified the principles for constructing instruments to measure within-subject change over time,[18-22] and it assumes that such instruments must be based on what patients feel is important. We provide two models: a Volkswagen model, for investigators with limited time and money who regard quality of life as a secondary outcome in their study, and a Rolls-Royce model, for those with substantial resources, an interest in questionnaire development and the belief that the treatment under investigation stands or falls on its effect on quality of life. We have used one or the other of these approaches in constructing questionnaires for clinical trials in patients with rheumatoid arthritis,[12] chronic airflow limitation,[23] chronic heart failure[23] and breast cancer.[15] We shall not attempt to discuss comprehensively the many associated methodologic controversies, but we will present a pragmatic approach to instrument development.

Steps in instrument development

The development of an evaluative instrument can be divided into six stages:[18] item selection, reduction of number of items, questionnaire format, pretesting, reproducibility and responsiveness, and validity (Table I).

Item selection

Items on the questionnaire must reflect areas that are important to patients suffering from the disease and therefore should be derived from what patients say about how the illness affects their lives. One can ask patients about areas of physical and emotional dysfunction and about the inconvenience and limitation that, from their point of view, arise from their illness. One problem with such direct questioning is that patients recall spontaneously only a small portion of all their areas of dysfunction, and not necessarily the most important ones. There are probably many reasons

Table I—Stages in development of two models for measuring quality of life

Stage	Rolls-Royce Model	Volkswagen Model
Item selection	Literature review Consultation (comprehensive) with health care workers Use of existing instruments Semistructured interview with 50 to 100 patients	Use of existing instruments Consultation with health care workers
Reduction of no. of items	Use of second item-selection questionnaire identifying item frequency and importance Choice of items with highest frequency–importance product or principal-component analysis	Only items needed in item-selection process selected
Questionnaire format	Choice of response-options scale: 7- to 10-point Likert or visual analogue scale Time specification Availability of previous responses to patients	As for Rolls-Royce model
Pretesting	Use of about 20 subjects Analysis of results to ensure that full range of response options is used	Use of two or three subjects
Sampling for above four stages	Use of random sample of patients eligible for subsequent trials to ensure representation of entire range of disease severity, age lifestyle etc.	Use of sample of convenience
Reproducibility and responsiveness	Questionnaire administration to stable patients, duplicating conditions of subsequent trial(s) Administration before and after intervention of known efficacy	No testing before trial(s)
Validity	Use of construct validity Use of *a priori* predictions	Use of face validity

for this, but one contributing factor may be patients' impression that health care professionals will not really be interested in the mundane effects of an illness on the patients' lives.

To determine the true frequency and importance of all items, one must provide a comprehensive series of probes that cover all possible areas of dysfunction. The nature of these probes will depend on the amount of detail the research study requires: questions may be general (e.g., on self-care), probe more specific functions (e.g., dressing) or even consider components of specific functions (e.g., doing up buttons).

The Rolls-Royce model dictates detailed semistructured interviews with 50 to 100 patients to determine all areas of dysfunction. We have found that few new items are volunteered or elicited when more than 100 patients are included. To determine the frequency and importance of each item, another sample of approximately 100 patients is questioned. In the second questionnaire patients are asked whether any of the items is a problem for them and the importance of each item. There are several approaches to determining item importance, each with its own theoretical and practical advantages and disadvantages. The easiest is to ask patients to rate the importance of each item that is a problem for them on a Likert scale (i.e., each item is rated from very important to not at all important). An alternative is to place item descriptions on cards and ask each patient to sort the cards into piles according to their importance.

Are two patient samples necessary? An alternative approach is to collect a comprehensive set of items before selecting a patient sample. One can poll content-area experts (physicians, nurses, physiotherapists etc.), use a small number of patients who are articulate or severely affected or both, and their spouses, review the relevant literature and examine questionnaires designed for the general population or for patients with related conditions. We have successfully used this approach in constructing our questionnaires for patients with heart and lung disease, women undergoing chemotherapy for breast cancer and patients with rheumatoid arthritis in a trial of occupational therapy.[15,23]

How should one choose patients for the item-selection questionnaire? With the Rolls-Royce model the patients should be a random sample of the group from which subjects for the controlled trial will subsequently be selected. This implies sampling of the complete spectrum of disease severity under consideration, and inclusion of patients from all subclasses (e.g., those of age, sex and duration of disease). The eligibility criteria for the trial should also be applied in selecting patients for the second questionnaire, which deals with item frequency and importance.

How many patients should be interviewed? The second item-selection questionnaire is designed to obtain a reasonable estimate of the frequency and importance of each area of dysfunction in the target population. With 100 patients the confidence interval for a frequency of 50% is about 10% (in other words, the true frequency would be between 40% and 60% in 95% of situations in which a frequency of 50% was obtained); with 50 patients the confidence interval is about 15%, and with 25 patients it rises to about 20%. The specifications of the Rolls-Royce model dictate the use of more than 50 patients.

The item-selection questionnaire is bypassed in the Volkswagen model. Here one reviews a couple of existing instruments, consults with one or two content-area experts and chooses items that one thinks appropriate.

How much does one lose in adopting the Volkswagen model for item selection? Our experience with the item-selection questionnaire is that one gets a few surprises but not many. We found, for example, that frustration and embarrassment because of coughing and the need to use medication in public were among the primary emotional problems of patients with chronic lung disease. Those with arthritis identified cosmetic effects, such as walking with a limp or having to use modified clothing, as important issues. Women with breast cancer felt that of the many unpleasant side effects of chemotherapy loss of hair was far more important than the others. They also felt that chemotherapy had very positive effects, because it gave them a sense of control over their lives. In retrospect, these findings may seem obvious, but we would not have predicted them either from our own experience or from published reports.

Reduction of number of items

Frequently the item-selection questionnaire will yield many more items than can be included in the final questionnaire. Important criteria for retaining items include the number of patients who listed the item as a problem (item frequency), the importance attached to the items and the potential responsiveness of the items (i.e., the item's ability to detect change if it is present). The main issue with the frequency and importance criteria is how they should be combined. Although a number of more sophisticated alternatives (such as factor analysis or principal-component analysis[24]) are favoured by some investigators, a simple and reasonable approach is to multiply the frequency of each item by its mean importance. Having done this one can then retain the items with the greatest frequency–importance product for the final questionnaire. However, there are other considerations.

Since the questionnaire is being designed for evaluative purposes (i.e., to measure within-subject change over time) there is no point in including items that are unlikely to demonstrate change with available intervention; doing so would compromise the instrument's responsiveness.[18] For example, while inability to continue at work may be a frequent and important problem for patients with

severe chronic lung disease, the available interventions are unlikely to permit a return to active employment; therefore, a question about work would be wasted in an evaluative questionnaire. Similarly, a question about difficulty in eating would not be appropriate if a quality-of-life measure were to be used in a trial in which most of the arthritic patients had fused upper limbs.

On the other hand, if one is intending to assess an intervention with specific goals, one would want to ensure that items related to those goals appear in the final questionnaire. For example, we wanted to use our questionnaire for patients with chronic lung disease to assess the results of a respiratory rehabilitation program.[23] Among the program's major goals were to decrease patients' fear and panic about dyspnea and to increase their sense of control over their lives. Therefore, we were inclined to choose items relating to these areas for the final questionnaire. Another example occurred in an ongoing study of the benefit of home occupational therapy in patients with rheumatoid arthritis. Here, substituting less painful activities or having some activities taken over by willing family members is an objective of therapy and so should contribute to a positive result. This contrasts with a trial of a disease-remitting agent, in which one would hope that patients would be able to undertake painful activities with less discomfort and to take on tasks formerly left to other family members.[12]

A final consideration in reducing the number of items is the way the items will be aggregated — that is, how their scores will be added to yield a final score, or scores, for each patient. Each dimension that one wishes to measure requires adequate representation on the questionnaire for two reasons: to decrease the variability in response found even in stable patients and to minimize the impact of idiosyncratic responses to individual questions. Therefore, it would not be appropriate to include only one question relating to emotional function; at least the three or four items with the highest frequency–importance product on the item-selection questionnaire should be included. The question of whether one should average across disparate dimensions such as emotional and physical function is a difficult one and will not be covered in this paper.

What if, after surveying the results of the item-selection questionnaire, one finds that restricting the number of items in the final instrument omits items that are very important to a substantial number of patients? We faced this problem in assessing activities associated with dyspnea in patients with chronic heart and lung disease[23] and disability in those with rheumatoid arthritis.[12] The activities were highly dependent on age, sex, disease severity and lifestyle, so that it was impossible to select a few core activities that would apply to most patients. In this situation one can use individualized questions. The strategy was developed by Scott and Huskisson[25] and is similar to that used in goal attainment scaling.[26,27] Patients are asked to choose their own items, which are retained for subsequent questionnaire administration. For example, we ask patients with chronic heart and lung disease to select five activities that are both important and frequent in their day-to-day lives but that cause exertional dyspnea. Then, on each subsequent visit they are asked about the severity of the dyspnea during the five activities. Asking patients to choose the five most important activities during which their joint disease limits them has also proved feasible in patients with rheumatoid arthritis.[12]

The number of questions remaining after the number of items has been reduced depends on how many other tests and questionnaires the patients will have to complete during the study. Generally we have aimed at keeping our questionnaire administration to less than 20 minutes and have found that 1 minute per question is a conservative estimate of the time required for all but the initial administration.

Reducing the number of items is not an issue with the Volkswagen model. During the initial item-selection process one simply selects the number of items one wishes to use.

Questionnaire format

If the content of the questionnaire overlaps with that of established instruments, wording can be borrowed from the latter. Guides for constructing questions are also available.[21,28,29] The questions must be specific about time: one cannot simply ask how the joint pain has been; one must ask how it has been during the last week, 2 weeks and so forth.

Response options refer to the categories or range that patients have in responding to questionnaire items. For example, if one wants to determine simply whether the patient has difficulty climbing stairs, only two response options — Yes and No — are necessary. However, when one wants to determine the degree of difficulty a wide variety of options must be available.

To ensure responsiveness to the questionnaire one must be able to detect small changes, if these changes occur, for each item. To ask about the presence or absence of difficulty climbing stairs inappropriately limits the patient to two response options; one could therefore not detect the effect of an intervention that reduced but did not eliminate difficulty climbing stairs. More suitable response options include a visual analogue scale (VAS) (which consists of lines, usually 100 mm in length, anchored by the extremes of the item being measured, on which patients mark their status for that item) and a Likert scale with multiple options (e.g., excellent, good, moderate, poor, very poor). Both options have their proponents,[30,31] but there is no evidence to support one over the other. Similarly, the optimal number of response options for a

Likert scale is unclear. On the basis of available data, selection of a VAS or a Likert scale with 7 to 10 response options is reasonable.

In the traditional approach to questionnaire administration patients have not been permitted to see their responses on previous occasions, so as to avoid bias — that is, a tendency to the same score even if change has occurred.[32] However, this view has recently been challenged.[33] We have found that showing patients their previous responses decreases the variability of responses from stable patients without attenuating changes in the questionnaire scores associated with responses to treatment.[34] We therefore recommend that the previous responses be made available to patients when evaluative questionnaires are being administered.

Pretesting

When questionnaires are first administered there are inevitably questions that some patients do not understand or that they find silly, inappropriate, embarrassing or confusing. Poor wording of questions or choice of response options may lead to patients' using only some of the response options (people hesitate to say they're feeling very depressed, for example). Pretesting is required to correct these problems. The Rolls-Royce model requires random selection and testing of another group of subjects eligible for the planned trial (once again ensuring representation of disease severity, age, lifestyle etc.). About 20 patients should be sufficient to identify major problems. Obvious problems are noted, and the interviewer reviews each question and asks subjects to explain what the question meant to them and why they chose a particular response option. Discrepancies between what was intended and what was understood are noted. Subjects are also asked to identify any questions that made them feel uncomfortable or embarrassed and to suggest how the questionnaire might be modified.

When the patients have been tested a statistical analysis is conducted. Responses to each question are examined to ensure that the full range of response options has been used. The questionnaire is then modified to eliminate ambiguities, delete questions that are offensive or beyond correction and ensure that the full range of response options is used for each question. When the modifications have been made the pretesting procedure is repeated.

With the Volkswagen model, pretesting involves choosing two or three subjects, administering the questionnaire and changing it only if obvious problems arise. In the construction and pretesting phases we have described, as well as in the instrument-testing phases that follow, most investigators will choose a strategy that falls somewhere between the Rolls-Royce and the Volkswagen models.

Reproducibility and responsiveness

Reproducibility has a number of synonyms, including reliability and precision, the point being that repeated administration of a questionnaire to stable patients should produce more or less the same results. The most commonly used approach to assessing reliability involves looking at the ratio of the variability between subjects to the total variability in responses (which includes variability attributable to both between-subject and within-subject differences). The resulting statistic is the Pearson's correlation coefficient, or a more sophisticated version, an intraclass correlation coefficient that takes into account systematic change in score over time.[35] These correlations tell us how good an instrument is at differentiating between subjects. However, with a questionnaire about quality of life we are interested in detecting change within subjects over time; thus, the magnitude of the variability between subjects is irrelevant, and the correlation coefficients may give misleading results.[36]

The usefulness of an evaluative questionnaire depends on its responsiveness — that is, its ability to detect clinically important changes even if the changes are small. Responsiveness is proportional to the change in score that constitutes a clinically important difference (the "signal" that the instrument is trying to detect) and inversely proportional to the variability in score in stable patients (the "noise", which makes the signal difficult to detect). The ratio of the minimal clinically important difference (or, if that is unavailable, the change produced by a treatment of known benefit) to the within-subject variability in stable patients is directly related to sample size requirements and can be used as an index of a questionnaire's responsiveness.[36]

Two studies are required to generate the data needed to determine questionnaire responsiveness: one to examine the variability in stable patients and the other to demonstrate that questionnaire scores change when real change has taken place. In the study designed to examine variability in stable patients the questionnaire is repeatedly administered to a group of patients who meet the eligibility criteria for the planned clinical trial and who are deemed stable by other criteria. The interval between administrations of the questionnaire, the number of times the questionnaire is given and the interval between first and final administration should duplicate what is planned in the clinical trial. The data from this study will yield an estimate of the variability in stable patients.

In the second study, designed to demonstrate that changes in the questionnaire score occur when real change has taken place, the questionnaire is administered to patients before and after application of an intervention of known efficacy. One may use either the same sample or a new group of patients, but once again the patients should meet the eligibility criteria for the subsequent trial. Ideally, the questionnaire score will demonstrate

not only improvement in quality of life but also sufficiently large improvement relative to the variability shown by stable patients. If there is no convenient therapy of known benefit, the situation becomes more difficult; one solution is to administer the questionnaire serially to patients in whom spontaneous improvement or spontaneous deterioration is expected.

The ratio between the change seen in patients in the second study to the variability in stable patients seen in the first study provides an estimate of questionnaire responsiveness. The larger the difference in questionnaire score in patients in whom there is a real change (the signal) the greater the responsiveness; the larger the difference in questionnaire score in patients who are really stable (the noise) the lower the responsiveness.

The Volkswagen model for assessing reproducibility and responsiveness is simple: begin the trial with the assumption that the index will prove equal to the task. One can then assess responsiveness in retrospect: variability in stable patients can be examined by reviewing results in a placebo group, and this variability can be compared with the treatment effect. This method will be satisfactory if the trial has a positive result. However, if the result is negative one cannot be sure that the treatment was ineffective or that the questionnaire was unresponsive. This is an important reason why an untested index should not be used as a primary measure of outcome.

Validity

An index is valid if it is measuring what it is supposed to measure. The simplest way of validating a questionnaire is to demonstrate that its results match a gold standard; however, a gold standard for quality of life is unavailable. Therefore, one must rely on "construct validity": Does the questionnaire behave in relation to other measures as one would expect if it was really measuring quality of life? Construct validity requires several predictions about how the results of the questionnaire should correlate with other related measures and then testing of these hypotheses. Because evaluative questionnaires are primarily concerned with measuring change, one must examine the correlations between change in the quality-of-life measure and change in other variables.[18]

For example, the questionnaire we have developed to measure the effects of quality of life in patients with chronic heart and lung disease examines both physical and emotional function.[23] At the time of questionnaire administration we ask the patient, a relative and the physician to make global ratings of the patient's physical and emotional function. If the questionnaire is really measuring quality of life we would expect the changes in ratings of physical function to bear a close relation to changes in the patient's global rating of physical function, a somewhat weaker relation to changes in the relative's rating of physical function and little, if any, relation to changes in the physician's rating of emotional function. Important additional validation comes from examining the results of physiologic measures. One would expect a strong correlation between changes in physical function and changes in exercise capacity and, to a lesser extent, changes in cardiac function or spirometric readings. These predictions can be tested by applying both the questionnaire and other measures of function in an open study before and after an intervention of known benefit.

The Volkswagen model relies on "face validity". If the questions in the index appear to be measuring quality of life, that is sufficient.

Conclusions

The importance of direct measurement of quality of life in establishing the benefit of treatment is increasingly becoming recognized.[22] To be confident that small but clinically important differences will not be missed requires a responsive measure. We have outlined an approach to the construction of disease-specific measures that can be applied to specific conditions or even specific trials once it has been determined that an adequate quality-of-life index is not already available. At each stage of instrument development the investigator can choose a pragmatic strategy that requires few resources or a sophisticated approach designed both to maximize the chances of constructing a useful instrument and to rigorously test its responsiveness and validity. If an investigator chooses the pragmatic route, efficiency is maximized, but the decreased confidence in the responsiveness and validity of the resulting index precludes use of the index as the primary measure of outcome in subsequent studies. While quality of life can be difficult to measure accurately, investigators should no longer shy away from including it as an outcome in studies designed to determine treatment benefit. The approaches we have outlined should make construction of a quality-of-life instrument for a specific trial less intimidating and improve the responsiveness and validity of the resulting measure.

This work was supported in part by the Ontario Ministry of Health and the St. Joseph's Hospital Foundation, Hamilton, Ont.

References

1. Ware JE, Brook RH, Davies-Avery A et al: *Conceptualization and Measurement of Health for Adults in the Health Insurance Study*, vol 1: *Model of Health and Methodology*, Rand Corp, Santa Monica, Calif, 1980
2. Kaplan RM, Bush JW, Berry CC: Health status: types of validity and the index of well-being. *Health Serv Res* 1976; 11: 478–507
3. Bergner M, Bobbit RA, Carter WB et al: The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* 1981; 19: 787–805

4. Parkerson GR Jr, Gehlbach SH, Wagner EH et al: The Duke-UNC Health Profile: an adult health status instrument for primary care. Idem: 806-828
5. Hunt SM, McKenna SP, McEwen J et al: A quantitative approach to perceived health status: a validation of study. *J Epidemiol Community Health* 1980; 34: 281-286
6. Spitzer WO, Dobson AJ, Hall J et al: Measuring the quality of life of cancer patients. *J Chronic Dis* 1981; 14: 585-597
7. Preistman TJ, Baum M: Evaluation of quality of life in patients receiving treatment for advanced breast cancer. *Lancet* 1976; 1: 899-901
8. Mor V, Laliberte L, Morris JN et al: The Karnofsky performance status scale. An examination of its reliability and validity in a research setting. *Cancer* 1984; 53: 2002-2007
9. Fries JF, Spitz PW, Young DY: The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. *J Rheumatol* 1982; 5: 789-793
10. Meenan RF: The AIMS approach to health care measurement: conceptual background and measurement properties. Idem: 785-788
11. Helewa A, Goldsmith CH, Smythe HA et al: Independent measurement of functional capacity in rheumatoid arthritis. Idem: 794-797
12. Tugwell P, Bombardier C, Buchanan W et al: The ability of the Mactar disability questionnaire to detect sensitivity to change in rheumatoid arthritis [abstr]. *Clin Res* 1983; 31: 239
13. Goldman L, Hashimoto D, Cook EF: Comparative reproducibility and validity of systems for assessing cardiovascular functional class: advantages of a new specific activity scale. *Circulation* 1981; 64: 1227-1234
14. Mahler DA, Weinberg DH, Wells CK et al: Measurement of dyspnea: description of two new indices, interobserver agreement and physiological correlations. *Am Rev Respir Dis* 1982; 24 (suppl 1): 138
15. Levine MN, Burns CM, Pirie KM et al: Quality of life in stage 2 breast carcinoma. *Clin Invest Med* 1984; 7 (suppl 2): 48
16. Alexander MR, Dull WL, Kasik JE: Treatment of chronic obstructive pulmonary disease with orally administered theophylline. *JAMA* 1980; 224: 2286-2290
17. Eaton ML, MacDonald FM, Church TR et al: Effects of theophylline on breathlessness and exercise tolerance in patients with chronic airflow obstruction. *Chest* 1982; 82: 538-542
18. Kirshner B, Guyatt GH: A methodological framework for assessing health indices. *J Chronic Dis* 1985; 38: 27-36
19. Bombardier C, Tugwell P: A methodological framework to develop and select indices for clinical trials: statistical and judgmental approaches. *J Rheumatol* 1982; 9: 753-757
20. Tugwell P, Bombardier C: A methodological framework for developing and selecting end points in clinical trials. Idem: 758-762
21. Bombardier C, Tugwell P: Measuring disability: guidelines for rheumatology studies. *J Rheumatol* 1983; 10: 68-73
22. Deyo RA: Measuring functional outcomes in therapeutic trials for chronic disease. *Controlled Clin Trials* 1984; 5: 223-240
23. Guyatt GH, Berman LB, Pugsley SO et al: Development of a responsive measure of quality of life for patients with chronic cardiorespiratory disease [abstr]. *Clin Res* 1984; 32: 222
24. Kleinbaum DG, Kupper LL: *Applied Regression Analysis and Other Multi-variable Methods*, Duxbury Pr, North Scituate, Mass, 1978: 376-414
25. Scott PJ, Huskisson EC: Measurement of functional capacity with visual analogue scales. *Rheumatol Rehabil* 1977; 16: 257-259
26. Kieresuk TJ, Sherman RE: Goal attainment scaling: a general method for evaluating comprehensive community mental health programs. *Community Ment Health J* 1968; 4: 443-453
27. Freeman HE, Sherwood CC: Research in large-scale intervention programs. *J Soc Issues* 1965; 21: 11-28
28. Woodward CA, Chambers LW: *Guide to Questionnaire Construction and Question Writing*, Can Public Health Assoc, Ottawa, 1983
29. Sudman S, Bradburn NM: *Asking Questions: a Practical Guide to Questionnaire Design*, Jossey-Bass, San Francisco, 1982
30. Huskisson EC: Measurement of pain. *Lancet* 1974; 2: 1127-1131
31. Downie WW, Leatham PA, Rhind VM et al: Studies with pain rating scales. *Ann Rheum Dis* 1978; 37: 378-381
32. Jacobsen M: The use of rating scales in clinical research. *Br J Psychiatry* 1965; 111: 545-546
33. Scott J, Huskisson EC: Accuracy of subjective measurements made with or without previous scores: an important source of error in serial measurement of subjective states. *Ann Rheum Dis* 1979; 38: 558-559
34. Guyatt GH, Townsend M, Taylor DW: Should patients see their previous scores? *J Chronic Dis* 1985; 38: 1003-1007
35. Kramer MS, Feinstein AR: Clinical biostatistics. LIV. The biostatistics of concordance. *Clin Pharmacol Ther* 1981; 29: 111-123
36. Guyatt GH, Walter SD, Norman G: Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* (in press)