The Future of Outcomes Measurement in Rheumatology

Robert M. Kaplan, PhD

Abstract

Quality-of-life (QOL) measurement has a rich history in rheumatology, and although the study of health measurements has expanded remarkably in recent years, there are a number of methodological issues that need attention. This article focuses on utility-based measures that are most relevant to public policy and some major issues that remain unresolved. The duration of a health condition, for example, is a central component of health outcome. However, the majority of measures currently used in rheumatology do not consider the duration of condition or the long-term consequences of disease or treatment. Similarly, there are many levels at which preference is expressed in the healthcare decision process, and despite a wealth of literature on QOL measurement, we still have much more to learn about how to present information to patients and how to use preference data that patients provide. Improvements in the utility assessment method are urgently needed as well, as different utility-based measures such as standard gamble, time trade-off, and rating scales are not comparable. This lack of comparability has important implications when estimating quality-adjusted life-years in cost-utility studies. Another key unresolved issue is whether the Medical Outcomes Study 36-Item Short Form and the Health Assessment Questionnaire offer appropriate data for cost-utility studies. The current estimation of clinical benefit is also controversial, as competing measurement tools offer different estimates of wellness at baseline. However, the different approaches yield surprisingly similar estimates of change. Clearly, this observation deserves further study and replication in future studies.

(Am J Manag Care. 2007;13:S252-S255)

The field of quality-of-life (QOL) measurement has a relatively brief history. The **Figure** summarizes the number of publications identified using the search term *quality of life* in PubMed between 1972 and 2007. There were no published papers on the topic in 1972, but the number of articles increased dramatically over the course of the last 35 years. By 2006, there were more than 6300 articles. Rheumatology has always been ahead of other specialties in the application of QOL measures. The Arthritis Impact Measurement Scales (AIMS)[1-3] and the Health Assessment Questionnaire (HAQ)[4-6] were among the first disease-specific measures. Crossing *arthritis* with *quality of life* in PubMed (July 2007) yields 1257 references.

Although the number of articles devoted to QOL measurement has expanded remarkably, important methodological issues still need attention. After considering the history of the field, this article will review some of those issues and identify important directions for future research. Because the field is so broad, we will focus on utility-based measures that are most relevant to public policy.

Neglected History

Many of the current measures borrowed structure and items from the work of JW Bush. Bush was a family physician who was concerned about universal healthcare coverage as early as the late 1960s. Working for the New York State Health Planning Commission, he set out to find the best way to allocate healthcare resources. The major obstacle to an equitable allocation scheme was the absence of a definition and measure of health outcome. In 1970, Fanshel and Bush published a profound blueprint for outcomes research.[7] Forty years ago, few people were thinking about ways to allocate resources, and the article was not widely recognized. Bush called for the systematic evaluation of health status using an index that separated health "states," or functional status, from "weights," which were qualitative judgments about the desirability of these conditions.[8] By 1973, Bush in collaboration with Patrick and Chen had developed and published a Health Status Index.[9,10] The index served as the basis for many contemporary measures, most notably the Quality of Well-Being Scale (QWB).[11-13] However, the basic components were used to create the measures for the RAND Health Insurance

Address correspondence to: Robert M. Kaplan, PhD, Professor and Chair, Department of Health Services, UCLA School of Public Health, PO Box 951772, Los Angeles, CA 90095-1772. E-mail: rmkaplan@ucla.edu.

Experiment, which later evolved into the Medical Outcomes 36-Item Short Form (SF-36).[14] The Health Status Index also served as the basis for the Functional Status Index,[15] the AIMS,[16] HAQ,[6] and several other measures.

Although there was some debate, Bush was probably the first to apply the concept of quality-adjusted life-years (QALYs). By 1971, long before the era of modern computing, Bush had applied Markov modeling for estimating health state transitions required to compute QALYs.[17] Markov models represent transitional processes by showing the probability of movement between defined states. Although these models are used in many fields, Bush was among the first to show how health status could be conceptualized as movement between defined health states over time. Believing that the term *quality* was too general for a health measure, Bush dropped references to QALYs and later used terms such as *discounted life-years*, *well-life expectancy*, and *well-years*.[18]

Before his untimely death in 1986, Bush laid out a series of methodological questions that needed attention. In the past 20 years, we have made remarkably little progress in addressing these issues, and it may be time to refocus our attention on some major issues that remain unresolved, such as those that follow.

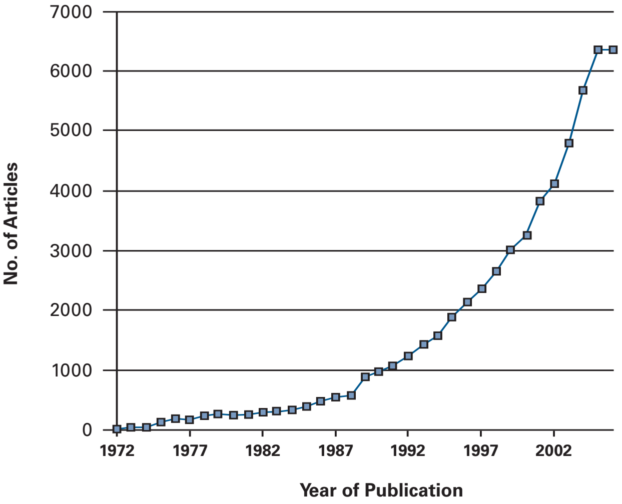
The Role of Duration and Prognosis in Health Outcome Assessment

The duration of a health condition is a central component of health outcome. Joint pain that lasts 1 month is not the same as pain that lasts 1 hour. In Bush's early formulation, changes in health status over time were represented by Markov chains of transition probabilities.[17] Today, some models of outcome explicitly include time. QALYs, for example, adjust survival time by QOL. However, the majority of measures do not explicitly consider duration of condition, or long-term consequences of disease or treatment. Consider, for example, a medication that offers short-term improvements in functioning but has serious long-term complications. Often the side effects are unanticipated and in a different organ system. A narrowly focused disease-targeted measure may miss adverse effects altogether. A comprehensive measurement system should offer an overall picture of health effects, and more work is needed on the measurement of life courses following treatment. This will require new approaches to simulation and modeling that merge results from clinical trials with long-term postmarketing evaluations.

The Measurement of Quality

There are many levels at which preference is expressed in the healthcare decision process. For example, a patient with rheumatoid arthritis (RA) may decide to cope with several undesirable symptoms to gain the potential benefit of a tumor necrosis factor (TNF)-alpha antagonist. The utility approach explicitly acknowledges that preferences are used to express the relative importance of various health outcomes. Whether we prefer back pain or an upset stomach caused by the anti-inflammatory drugs used to treat the pain is a value judgment. Not all symptoms are of equal importance. Most patients would prefer mild fatigue (a side effect of treatment) to a severe pain (the symptom eradicated by treatment). Outcome models implicitly include these judgments. We still know little about how patients trade off symptoms and risks. Despite rich literature on QOL measurement, we still have much more to learn about how to present information to patients[19] and how to use preference data that patients provide.[20]

■ **Figure.** Articles by Search Term *Quality of Life* and Year of Publication



Utility Assessment Method

Nearly 35 years ago, Patrick et al reported that different methods for measuring utility did not give the same results.[10] Publications on the noncomparability of the Standard Gamble (SG), Time to Trade-off (TTO), and Rating Scales (RS) continue to proliferate. In July of 2007, PubMed identified 632 articles on SG and 132 articles under the combination of SG and TTO. All of the methods are designed to place level of wellness on a continuum anchored by 0.0 for death and 1.0 for perfect health. We would expect different methods to offer comparable numbers along this 0.0 to 1.0 scale. However, it is common to find that RS, SG, and TTO methods yield different utilities. For example, Khanna and colleagues[21] reported that patients with systemic sclerosis gave ratings of 0.62 using RS, 0.83 on the SG, and 0.88 using TTO. These differences may have important implications for estimating QALYs.

Although these findings are important, we do not need additional studies showing SG, TTO, and RS yield different numbers; we already know that. It should not be a surprise that methods based on different conceptual models and scales that ask different types of questions yield different results. In the future, we need more investigation of which methodology is correct. Economists prefer SG because it is linked to economic theory. On the other hand, the evidence supporting the theory may not be as strong as is believed.[22,23] Empirical tests of measurement properties have been developed, and some of the assumptions behind the measurement models can be tested empirically.[24] Future research must test the assumptions behind the measurement models.

Using the SF-36 or HAQ for Cost-utility Studies

We previously believed that the application of utility-based measures was necessary to estimate QALYs. Yet most clinical trials use the SF-36 or HAQ Disability Index instead of one of the utility-based measures. Over the past few years, a variety of methods have been developed for estimating QALYs using SF-36 data. The first method was described by Fryback and colleagues using data from the Beaver Dam Health Outcomes Study.[25] A regression model was used to translate SF-36 components into QWB scores. A similar method by Nichol et al[26] was developed to translate SF-36 scores into the Health Utilities Index Mark 2 (HUI2). Brazier et al[27] have developed a method known as the SF-6D that uses direct utility ratings of SF-36 components. There have been attempts to develop QALY measures that are specific to arthritis.[28] However, it is not clear this approach is valuable for policy analysis. The purpose of QALYs is to obtain generic outcome units that can be used to compare the cost-effectiveness of very different interventions (ie, disease-modifying drugs in RA vs lung volume surgery in chronic obstructive pulmonary disease). To accomplish these broad comparisons, health outcomes should be measured in a generic sense. The development of an RA-QALY would result in a unit that cannot be used for these broad comparisons.

The new methods for imputing utility scores from the SF-36 constitute a methodological advance. However, there have been few evaluations of these measures against clinical outcomes or against other validated QOL measures. Successful estimation of QALYs from SF-36 data has the potential to bring together the psychometric and the decision-theory–based approaches. After nearly 30 years of separation, this is an encouraging development.

Estimation of Clinical Benefit

Investigators are fond of creating tables showing the cost per QALY produced by different interventions. However, if the different interventions are evaluated with nonequivalent methods, it makes little sense to place all of the values in the same table. One way to think of different utility-weighing methods is that they are different scoring systems for the same health states. It is possible to use different sets of weights (ie, SG, TTO, RS) attached to the same health outcome observations. In other words, the same patients complete a health status questionnaire, and then different weights can be applied to determine if the conclusions are affected by the utility weighting system.

We compared 2 versions of the HUI, the EQ-5D and the Brazier method for measuring outcome among patients with RA. Scores produced by the different indices varied considerably. All measures use the same 0.0 to 1.0 continuum, but the mean baseline scores produced by the different methods ranged from 0.44 to 0.81. If the measures are offering such different values, can the scores be trusted? As part of this analysis, we also evaluated the capability of each measure to detect significant clinical change. Each patient had participated in a clinical trial of a TNF-alpha antagonist. On the basis of clinical evaluations, they were categorized into 1 of 3 groups. One group failed to achieve a 20% improvement according to criteria from the American College of Rheumatology (ACR20). A second group achieved an ACR20 but not a 50% improvement (ACR50), and a third group achieved an ACR50. For each measure, we estimated the sensitivity for detection of clinical change using the eta^2 statistic that describes the percentage of variance explained. The different outcome measures were remarkably similar in their sensitivity to clinical change. In each case, the measure was highly responsive to change, and the eta^2 values were quite comparable across measures.[29]

These findings suggest that the measures offer quite different estimates of wellness at baseline. However, cost-utility analysis is based on measures of change. The different approaches yield surprisingly similar estimates of change. This implies that the selection of a measure may not have a profound impact on the QALY estimates used for policy analysis. Clearly, this observation deserves further study and replication in future studies.

QOL measurement has a rich history in rheumatology. However, many of the methodological issues raised during the birth of the field remain unresolved today. There is an important future for systematic methodological research on health outcomes measurement.

REFERENCES

1. Meenan RF. The AIMS approach to health status measurement: conceptual background and measurement properties. *J Rheumatol.* 1982;9:785-788.

2. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis. The Arthritis Impact Measurement Scales. *Arthritis Rheum.* 1980;23:146-152.

3. Meenan RF, Yelin EH, Nevitt M, Epstein WV. The impact of chronic disease: a sociomedical profile of rheumatoid arthritis. *Arthritis Rheum.* 1981;24:544-549.

4. Fries JF, Koop CE, Beadle CE, et al. Reducing health care costs by reducing the need and demand for medical services. The Health Project Consortium. *N Engl J Med.* 1993;329:321-325.

5. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. *Arthritis Rheum.* 1980;23:137-145.

6. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the Health Assessment Questionnaire, disability and pain scales. *J Rheumatol.* 1982;9:789-793.

7. Fanshel S, Bush JW. A health-status index and its applications to health-services outcomes. *Oper Res.* 1970;18:1021-1066.

8. Bush JW, Chen MM, Patrick DL. Cost-effectiveness using a health status index: Analysis of the New York State PKU screening program. In: Berg R, ed. *Health Status Indexes.* Chicago, IL: Hospital Research and Education Trust; 1973:172-208.

9. Patrick DL, Bush JW, Chen MM. Toward an operational definition of health. *J Health Soc Behav.* 1973;14:6-23.

10. Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Serv Res.* 1973;8:228-245.

11. Kaplan RM, Alcaraz JE, Anderson JP, Weisman M. Quality-adjusted life years lost to arthritis: effects of gender, race, and social class. *Arthritis Care Res.* 1996;9:473-482.

12. Kaplan RM, Anderson JP, Patterson TL, et al. Validity of the Quality of Well Being Scale for persons with human immunodeficiency virus infection. HNRC Group. HIV Neurobehavioral Research Center. *Psychosom Med.* 1995;57:138-147.

13. Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. *Health Serv Res.* 1976;11:478-507.

14. Stewart A, Ware JE Jr, Brook RH. The meaning of health: understanding functional limitations. *Med Care.* 1977;15:939-952.

15. Reynolds W, Rushing W, Miles D. The validation of a function status index. *J Health Soc Behav.* 1974;15:271.

16. Meenan RF. The AIMS approach to health status measurement: conceptual background and measurement properties. *J Rheumatol.* 1982;9:785-788.

17. Bush JW, Zaremba J. Estimating health program outcomes using a Markov equilibrium analysis of disease development. *Am J Public Health.* 1971;61:2362-2375.

18. Kaplan RM, Bush JW. Health-related quality of life for evaluation research and policy analysis. *Health Psychol.* 1982;1:621-680.

19. Zikmund-Fisher BJ, Fagerlin A, Ubel PA. Mortality versus survival graphs: improving temporal consistency in perceptions of treatment effectiveness. *Patient Educ Couns.* 2007;66:100-107.

20. Damschroder LJ, Roberts TR, Zikmund-Fisher BJ, Ubel PA. Why people refuse to make tradeoffs in person tradeoff elicitations: A matter of perspective? *Med Decis Making.* 2007;27:266-280.

21. Khanna D, Ahmed M, Furst DE, et al. Health values of patients with systemic sclerosis. *Arthritis Rheum.* 2007;57:86-93.

22. Broome L. The goal is quality improvement. *Nurs Manage.* 1993;24:51-52.

23. Richardson J. Cost utility analysis: what should be measured? *Soc Sci Med.* 1994;39:7-21.

24. Zhu S, Anderson N. Self-estimation of weight parameter in multiattribute analysis. *Organ Behav Hum Decis Process.* 1991;48:36-54.

25. Fryback DG, Lawrence WF, Martin PA, Klein R, Klein BE. Predicting quality of well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study. *Med Decis Making.* 1997;17:1-9.

26. Nichol MB, Sengupta N, Globe DR. Evaluating quality-adjusted life years: estimation of the Health Utility Index (HUI2) from the SF-36. *Med Decis Making.* 2001;21:105-112.

27. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ.* 2002; 21:271-292.

28. Chiou CF, Suarez-Almazor ME, Sherbourne CD, et al. Development and validation of a preference weight multiattribute health outcome measure for rheumatoid arthritis. *J Rheumatol.* 2006; 33:2409-2411.

29. Kaplan RM, Groessl EJ, Sengupta N, Sieber WJ, Ganiats TG. Comparison of measured utility scores and imputed scores from the SF-36 in patients with rheumatoid arthritis. *Med Care.* 2005;43:79-87.

Notes