OUTCOMES MANAGEMENT
AND
RESOURCE ALLOCATION:

HOW SHOULD
QUALITY OF LIFE BE MEASURED?

D.C. Hadorn

HPRU 93:7D JULY, 1993

OUTCOMES MANAGEMENT AND RESOURCE ALLOCATION:

HOW SHOULD QUALITY OF LIFE BE MEASURED?

David C. Hadorn, M.D., M.A.

HPRU 93:7D JULY, 1993

HEALTH POLICY RESEARCH UNIT

CENTRE FOR HEALTH SERVICES AND POLICY RESEARCH
#429 - 2194 HEALTH SCIENCES MALL
UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, B.C. CANADA
V6T 1Z3

The Centre for Health Services and Policy Research was established by the Board of Governors of the University of British Columbia in December 1990. It was officially opened in July 1991. The Centre's primary objective is to co-ordinate, facilitate, and undertake multidisciplinary research in the areas of health policy, health services research, population health, and health human resources. It brings together researchers in a variety of disciplines who are committed to a multidisciplinary approach to research, and to promoting wide dissemination and discussion of research results, in these areas. The Centre aims to contribute to the improvement of population health by being responsive to the research needs of those responsible for health policy. To this end, it provides a research resource for graduate students; develops and facilitates access to health and health care databases; sponsors seminars, workshops, conferences and policy consultations; and distributes Discussion papers, Research Reports and publication reprints resulting from the research programs of Centre faculty.

The Centre's Health Policy Research Unit Discussion Paper series provides a vehicle for the circulation of preliminary (pre-publication) work of Centre Faculty and associates. It is intended to promote discussion and to elicit comments and suggestions that might be incorporated within the work prior to publication. While the Centre prints and distributes these papers for this purpose, the views in the papers are those of the author(s).

A complete list of available Health Policy Research Unit Discussion Papers and Reprints, along with an address to which requests for copies should be sent, appears at the back of each paper.

TABLE OF CONTENTS

ABSTRACT

The relentless increase in health care expenses, coupled with persistent concerns about the quality and appropriateness of medical services, has brought increasing pressure to bear on researchers to develop more efficient strategies for determining "what works" in medicine. Several years ago, Paul Ellwood recommended that society arrange to regularly collect information concerning the health-related quality of life (HRQOL) of patients with medical problems and conditions. Coupled with demographic, clinical, and treatment data, this HRQOL outcome information would permit researchers to determine which services provide significant benefit to which types of patients.

For many reasons, Ellwood's vision of large-scale "outcomes management" programs has not come to pass. Probably the most significant impediment has been the absence of a very brief, generic HRQOL survey instrument which is calibrated according to empirically derived values and preferences. This discussion paper describes an effort to create and test such a questionnaire.

In Part A we describe the process of calibrating a new, four-item HRQOL questionnaire based on preferences elicited (in various ways) from about 600 people from different walks of life. During this process we observed few systematic differences in preferences across demographic lines; moreover, people with medical conditions or disabilities did not rate HRQOL-problem health states much differently than people without those conditions or disabilities.

Part B describes our clinical testing of the mail-in questionnaire in a cohort of 400 cancer patients. The questionnaire was administered three times to each patient over a six month period. We found that the questionnaire was able to capture patients' "true" HRQOL, as determined by subsequent careful, standardized (and blinded) telephone interviews. We conclude that use of our four-item questionnaire -- or of its global item alone -- can provide valid information concerning patients' HRQOL.

Part C discusses several key issues pertaining to the interpretation of observational HRQOL outcome data. To strengthen causal inferences drawn from these data, we recommend that the definition of HRQOL be restricted and standardized, and that the purpose of collecting outcome data be carefully explained as part of a public education program. Reporting one's HRQOL outcomes should come to be seen as a civic duty, like voting. Large sample sizes will be needed to control for potential confounding factors that could complicate (or preclude) inferences concerning treatment effectiveness. Finally, we discuss the central problem of outcome research: identifying "types of patients" who either do or do not derive significant benefit from specified treatments and procedures.

We conclude that very brief, even single-item, questionnaires can be appropriate vehicles for obtaining valid data concerning patients' HRQOL outcomes, and that valid inferences can be drawn from these data concerning the effectiveness of medical interventions. Large-scale outcomes management studies can provide an efficient and powerful strategy for determining "what works" in medicine (or, more precisely, "what works for whom"). By identifying when services have or have not been shown to provide significant net health benefit, society can ensure that its health care resources are allocated wisely.

LIST OF FIGURES

Part A

Calibration of a Brief Questionnaire and a Search for Preference Subgroups

I. Background and Significance

Five years have passed since Paul Ellwood first introduced the term "outcomes management" to the health policy debate.[1] Ellwood envisioned the creation of a "permanent national medical data base that uses a common set of definitions for measuring quality of life." Patients would regularly supply information concerning their health-related quality of life as part of a large-scale quasi-experiment designed to determine the outcomes of care and the effectiveness of medical and surgical services. Responses to the outcome questionnaires would be linked to medical records containing information about patients' conditions and treatments. This activity, Ellwood believed, would provide a "central nervous system" for a health care system increasingly characterized by "uninformed patients, skeptical payers, frustrated physicians, and besieged health care executives."

Ellwood's vision helped launch what is now commonly referred to as the "outcomes movement."[2] In an editorial published a few months after Ellwood's article, Arnold Relman heralded the dawn of the "third revolution in health care, the Era of Assessment and Accountability."[3] Relman endorsed the idea of "linking medical management decisions to new, systematic information about outcomes," saying this process would "improve the quality and effectiveness of health care and provide a much firmer base for future economic decisions." This apparent allusion to resource allocation decision-making was echoed more explicitly a couple of years later by John Wennberg,[4] who asserted that the use of outcome information might forestall the need for society to ration truly effective services by "sort[ing] out what works in medicine."

This bright promise seems to have dimmed somewhat of late. Indeed, despite sustained intensive efforts by many talented researchers and policy analysts, we are not much closer to knowing how, exactly, society might make use of outcomes information to set priorities within the health care system.

What's the Problem?

The basic idea sounds simple enough. First, determine the health outcomes associated with different treatments. Next, determine how people feel about those outcomes.

Finally, give priority to treatments that produce more-preferred outcomes. Such an outcome-based approach to resource allocation, although potentially vulnerable to charges of discrimination,[5] [6] is eminently reasonable, fair, and, indeed, necessary.

It is anything but simple, however. To date, only one instance of an outcome- and preference-based effort to set priorities has occurred: the trailblazing (and highly controversial) work of the Oregon Health Services Commission.[6] [7] [8] Unfortunately, although the Oregon project provided a wealth of experience on one possible approach to estimating and dovetailing outcomes with preferences, the result of that project -- a priority list containing 688 condition-treatment pairs -- is of questionable utility. Because of the wide range of procedures and indications contained within each "line item" on the list, substantial additional specification will be needed before the list can be applied to actual patients.[6]

The Oregon experience notwithstanding, the real-world use of outcomes to set health care priorities remains a seemingly distant goal. Moreover, systematic outcomes management, as envisioned by Ellwood, is essentially no closer to implementation than when it was first proposed. Why is this? There are many reasons, including (1) lack of a standard electronic medical record for entering and retrieving the necessary data;[9] (2) physician resistance to the use of formal patient outcome questionnaires;[10] [11] and (3) reluctance to base policy on causal inferences drawn from non-experimental data.[12] The first two of these problems can be overcome with greater federal leadership; the third will require, in addition, careful methodological consideration.[13]

None of the above factors is the most fundamental obstacle to creation of a large-scale system of outcomes management, however. A more basic problem is the lack of a standard, very brief, generic quality-of-life outcome questionnaire that is *calibrated according to empirically derived public preferences.* Creation of such a questionnaire constitutes the essential first step toward development of an effective system of outcomes management.

Need for a Standard, Brief, Generic Outcome Questionnaire

Clinical studies of particular conditions and treatments often use as endpoints condition- and treatment-specific outcomes (e.g., shortness of breath or nausea caused by chemotherapy). Such outcome measures are not suitable for purposes of resource allocation, however, as they do not permit comparison of different treatments and procedures.[14] [15] [16]

> How important, for example, is a coronary bypass operation for a specified clinical condition vs. chemotherapy for a particular type of cancer? If money is

> tight, which service should be funded if both cannot be? Answers to this sort of question will depend on the comparability of outcome data, which is possible only through the use of generic measures.[17,(p.775)]

James Bush and his colleagues were among the first to recognize the importance of generic measures (e.g., pain or physical suffering) for resource allocation decision-making. Generic measures formed (and continue to form) an integral part of the quality-adjusted life year concept introduced by these investigators over 20 years ago.[18] [19] More recently, RAND researchers developed and used generic measures in the Medical Outcome Study in order "to compare patients who have different conditions by providing a common yardstick"[20] against which to measure outcomes.

Recognition of these advantages has led to the development of a host of generic outcome questionnaires,[21] the most commonly used of which today is known as the SF-36.[22] Based on the outcome measures used in the Medical Outcome Study,[20] the SF-36 (for Short Form--36 items) is now in widespread use throughout the United States, including many of the Patient Outcome Research Teams sponsored by the federal Agency for Health Care Policy and Research.[23] The SF-36 contains items related to several aspects of health-related quality of life (HRQOL), including pain, extent of limits in daily activities, mental health, and energy level. Researchers plan to use the input obtained from the SF-36 to assess the HRQOL outcomes associated with various treatments.

Questionnaire Design and Calibration

In assessing the importance of the HRQOL outcomes reported by patients in a system of outcomes management, scoring of the questionnaires used to detect and measure these outcomes must reflect the actual values or preferences of the public.[15] Unfortunately, like many other generic questionnaires, the SF-36 is constructed without explicit reference to the relative priority people place across and within different dimensions of HRQOL. For example, the SF-36 contains fourteen items pertaining to physical functioning and performance of daily activities, but only two items related to pain. Unless patients value functioning seven times as much as relief or avoidance of pain (an empirical question addressed in this study), the results from the SF-36 will be biased toward functioning and away from pain if aggregate scores are calculated.

Similarly, scoring within items on the SF-36 is essentially arbitrary. For example, all four of the following transitions are assigned one point:

1. A change from "good" to "fair" health

2. A change from "limited a little" to "limited a lot" in bathing and dressing one's self

3. A change from "feeling calm and peaceful "most of the time" to "a good bit of the time"

4. A change from "accomplishing as much as one would like" to accomplishing less than one would like (because of physical or emotional problems).

It is possible, however, that most people might consider transition #4 to be substantially worse than #3, or vice versa. Similar instances of unequal weighting of preferences might be found for many or most of the other items.

The lack of empirical preference weights for the items contained in the SF-36 (and most other HRQOL questionnaires) is a significant shortcoming, especially if these instruments are to be used for the socially sensitive purpose of setting health care priorities. By contrast, three other well-known generic instruments are calibrated according to empirically derived preferences: the Quality of Well-Being Scale (QWB),[24] the Nottingham Health Profile (NHP)[25] [26], and the Sickness Impact Profile (SIP).[27]
Unfortunately, all of these questionnaires are probably too long to be used in large-scale patient outcome studies, Under a system of outcomes management, patients must continue to complete the health status surveys regularly after they have contacted the health care system, perhaps every six to twelve months following initial contact. This means, in turn, that the surveys will need to be mailed to patients' homes at regular intervals for self-administration and returned to a central receiving center. Very high response rates will be required if these outcome data are to command enough respect to guide priority-setting. From the perspective of cost and ease of response, a coded postal-card version of the questionnaire would probably be ideal for this purpose; perhaps three or four items would be the maximum for such a format. (See Part C for further discussion of survey logistics.)

The Rosser-Kind Index is the closest extant example of such an instrument.[28] This questionnaire has been used primarily for resource planning purposes in the United Kingdom and Europe, but has seen little use in North America. The Rosser-Kind Index consists of two items: one concerning pain (none, mild, moderate, and severe), the other on physical functioning (eight levels, ranging from no limits to unconscious). We believe that these generic HRQOL dimensions represent a reasonable distillation of the range of symptoms and disabilities caused by medical conditions, diseases, and treatments. (See Reference [17] for further discussion of this issue.)

Based on our experiences in previous studies,[29] [30] we modified the Rosser-Kind Index to produce a four-item questionnaire (Figure 1) for use in outcomes management and resource allocation. In Part A of this discussion paper we report the results of the questionnaire calibration phase of this project. In Part B we describe our experience using the questionnaire in a cohort of 400 cancer patients.

Preference Subgroups and the Problem of Discrimination

During the process of calibrating our test questionnaire -- dubbed the Quality of Life and Health Questionnaire -- we also conducted a search for coherent preference subgroups. An important unresolved issue in the field of resource allocation is whether people's preferences differ significantly based on demographic characteristics or, particularly, on whether they have experienced (or are experiencing) medical conditions or disabilities. Concern over such differences was the stated reason for the initial denial of the waiver needed by the State of Oregon to implement its much-discussed effort to set health care priorities in its Medicaid program.[31]

What is the evidence for differing preferences? There are some data suggesting that preferences for *specific health states* may vary based on experience with those states. People who have experienced specific states of poor health may rate those states higher,[32] [33] [34] lower,[35] or the same[36] [37] as people without comparable experience. In contrast to these inconsistent findings, people's preferences for *generic* outcomes are remarkably consistent, irrespective of demographic or clinical characteristics.[24] [38] [39] [40] [41] [42] [43] Indeed, this finding is perhaps the most consistent result within the entire field of preference measurement. One might speculate that, in addition to their advantages in the outcome-assessment arena (as discussed above), generic outcomes may be easier for people to comprehend (and to rate) in terms of preferences. This could be true because, although most people have not experienced any given specific health state, almost everyone has experienced pain and at least temporary limits on daily activities. This experience might facilitate the assignment of preference values to generic health states.

The evidence concerning the uniformity of preferences for generic outcomes has at least two shortcomings: (1) the small numbers of people enrolled in these studies (typically less than 200) and (2) the lack of systematic searches for coherent preference subgroups. In the present study, we obtained preferences from a diverse sample of 599 people, and made a systematic effort to detect subgroups of people with significantly different preferences.

II. METHODS

Sample

Subjects were 618 individuals recruited throughout the State of New Jersey from a wide variety of settings, including church and civic organizations and support groups for patients with chronic illnesses or conditions. A special effort was made to include subjects who represented a wide cross-section of demographic and clinical variables, including people with physical disabilities. The questionnaire was administered in small groups of about 10-12 people each. A trained facilitator provided preliminary instructions and assistance as necessary. The completion rate was high overall; out of 618 surveys administered, 599 (96.9%) were completed.

Respondents were 61.4% female. The majority were Caucasian (70.6%) and African-American (18.3%), with the remainder (11.1%) mainly of Hispanic and Asian ethnicity. Ages ranged from 14 to 91, with a mean of 48.9, a standard deviation of 20, and 90% of cases between the ages of 20 and 75.

The sample was somewhat skewed in the direction of higher education and above-average income. For example, 21.4% of respondents reported an advanced degree, and 37.1% reported yearly incomes of above $50,000. Lower-income and less well educated subjects were, however, reasonably represented; for example, 8.6% of subjects reported not having completed high school and 20.3% reported yearly incomes of less than $15,000.

Preference-Measurement Instrument

Based on the results of two pilot studies,[29 30] we developed a specially designed instrument designed to calibrate the four-item questionnaire shown in Figure 1. This latter questionnaire was administered to a cohort of cancer patients, as described in Part B of this discussion paper. The longer questionnaire used during this part of the study was designed to assign weights to the 16 health states defined by the four levels each of physical suffering and activity limitations. For reasons discussed more fully elsewhere[17] we analyzed separately the level of emotions and of overall quality of life reported by patients, as a sort of "baseline adjustment" in patients' response level. (See Part B.)

Each participant received a 32-page health preference survey that consisted of several sections. The first section of the questionnaire asked subjects to rate the 16 health scenarios on a scale of 0 = worst possible quality of life to 10 = best possible quality of life. Scenarios were constructed by pairing each of four level of physical suffering (none, mild, moderate, severe) with each of four levels of activity limitations (none, mild, moderate, severe). Accompanying each scenario was a cartoon figure that represented the combination of physical suffering and activity impairment. (For illustrations and a more extensive discussion of the cartoons and rating formats and their validation, see Ref. [30].) The scenarios were presented individually in random order. We refer herein to these sixteen ratings as the direct ratings.

The second section contained pairwise comparisons of health scenarios. The scenarios compared were taken from the 16 scenarios above. In this section, a pair of scenarios was presented and the subject asked to rate the extent to which one was preferred to the other. For example, the first item in this section asked subjects to compare condition 1 = (no physical suffering + mildly limited activities) with condition 2 = (mild physical suffering + no activity limitations). Note how this choice entails a trade-off between physical suffering and activity limitation; all the pairwise comparisons had this characteristic.

We constructed 13 comparisons so that each of the 16 original scenarios, except the (no suffering + no limitations) and (severe suffering + severe limitations) scenarios were compared with at least one other scenario. The two extreme scenarios were excluded because in pilot studies the former was always strongly preferred to all other conditions and all other conditions were strongly preferred to the latter. Ratings were made on a scale of -5 = strongly prefer condition "1" to +5 = strongly prefer condition "2", with 0 = no preference. These are the paired comparison ratings.[30 44]

Section 3 of the questionnaire investigated how subjects valued tradeoffs between increased life expectancy brought about by medical treatment and concomitant reduced quality of life. The format was similar to the paired comparison items, except that each scenario included a specific life expectancy. For example, the first item in this section asked subjects to compare treatment outcome 1 = (no physical suffering + mildly limited activities + 12 month life expectancy) with treatment outcome 2 = (mild physical suffering + mildly limited activities + 13 month life expectancy). The section contained 12 items with this format. Ratings were made on the same scale of -5 to 5 as with the paired comparison ratings. We refer to these ratings as the time-tradeoff ratings.

Section 4 of the questionnaire examined how subjects believed their emotional status and outlook on life are affect their quality of life. We do not report results from this section here.

Section 5 contained four items that attempted to ascertain subjects' current quality of life (QOL). The first item asked subjects to rate their level of physical suffering (none, mild, moderate, or severe) during the previous four weeks. The second asked them to rate their activity limitations (none, mild, moderate, or severe) over the same period. The third item asked subjects to rate their outlook on life (good, fair, somewhat poor, or very poor) over the previous four weeks. The last item asked subjects to report their current overall quality of life on a scale of 0 = worst possible to 10 = best possible. Note that this section is similar to the actual quality-of-life survey tested in a cohort of cancer patients (Part B). Other sections asked about the presence or absence of specific health conditions and demographic data, including the presence of disabling conditions.

Statistical Analysis

As described above, subjects assigned direct ratings (on a 0 to 10 scale) to each of the sixteen health states formed by crossing four levels of physical suffering with four levels of activity limitation. The means of these direct ratings constituted the basic calibration of the Quality of Life and Health Questionnaire (Figure 1). To determine whether one health state was preferred over another, we assessed differences in ratings across pairs of health states using paired Student's t-tests, with statistical significance set at $p < .01$. We considered differences in ratings between pairs of states to be *clinically* significant if one state was rated at least one-half point higher than the other state on our eleven-point scale.

As a validity check on the direct ratings, we calculated scale values for the same sixteen health states using the paired-comparison ratings.[30] For this analysis we used the PAIR program for graded paired-comparisons.[30] We then calculated the Pearson product-moment correlations between the scale values obtained from direct ratings and those derived from the paired-comparisons. The time-tradeoff ratings were not compared with the direct ratings or paired comparisons, but were used only in testing for the existence of coherent preference subgroups.

In searching for differences in preferences across subgroups, we began with a multivariate statistical test of association (forward stepwise regression) between the demographic or clinical variable and each item set (the direct ratings, the paired comparison ratings, and the time-tradeoff ratings). If a significant association between a demographic variable and an item set emerged, we then examined the association of the demographic variable with each item in the item set individually. All statistical analyses were conducted both with and without statistical adjustment (via partial correlations) for potential confounding variables. Statistical significance was defined as $p < .05$, although in many cases p values were much lower.

FIGURE 1

THE QUALITY OF LIFE AND HEALTH QUESTIONNAIRE

Instructions: Please check the box next to the word ([x]) that best describes how you have been feeling over the past week.

<u>PHYSICAL SUFFERING:</u> Headaches, chest or back pain, arthritis, nausea or vomiting, shortness of breath, dizziness, itching, etc.

NONE [ ] Physical suffering is rarely or never a problem.

MILD [ ] Somewhat bothersome problem but generally goes away by itself.

MODERATE [ ] More troubling problem with suffering.

SEVERE [ ] Extremely disturbing problem with suffering.

<u>EMOTIONS/OUTLOOK ON LIFE:</u> Feeling happy or sad, peaceful or nervous, and how much you look forward to getting up in the morning. How much of a problem:

NONE [ ] Emotions and outlook on life are rarely or never a problem.

MILD [ ] Somewhat bothersome problem with feeling downhearted and blue.

MODERATE [ ] More troubling problem with feeling depressed or nervous.

SEVERE [ ] Extremely disturbing problem with feeling depressed or nervous.

<u>DAILY ACTIVITIES:</u> Working or favorite pastimes, doing things with friends and family, and basic self-care activities -- such as: bathing, getting dressed, eating, and going to the bathroom. How much of a problem:

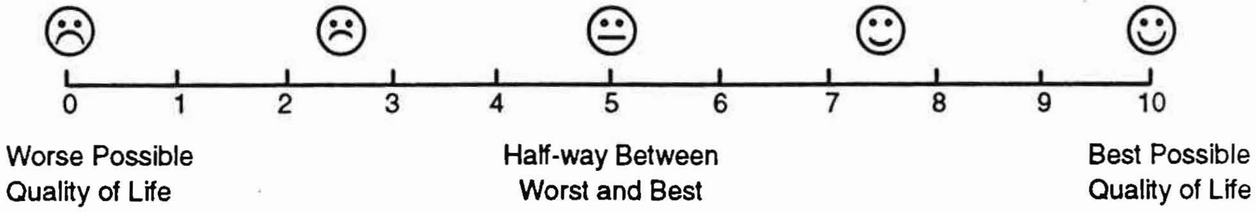
NONE [ ] Daily activities are rarely or never a problem.

MILD [ ] Somewhat bothersome problem with being limited in activities.

MODERATE [ ] More troubling problem with having to reduce activities.

SEVERE [ ] Extremely disturbing problem with having to reduce activities.

OVERALL, HOW WOULD YOU RATE YOUR QUALITY OF LIFE?
(Circle one number)



III. RESULTS

The mean ratings and standard deviations for the sixteen basic health states are depicted in Figure 2. As a rule, differences in ratings of 0.3 units or greater were statistically significant ($p < .01$). The correlation between the direct rating values and the scale values derived from the paired-comparisons was $|r| = .94$.

Effect of Subjects' Quality of Life on Ratings

The first factor we considered in our search for differences in preference patterns was subjects' perceived level of their own quality of life. For example, persons who are very healthy might rate an average health state as undesirable, whereas subjects who have a poor current health-related quality of life might rate the same state as desirable. As noted earlier we used four items to measure subjects' current HRQOL: current physical suffering, current activity limitations, current emotional outlook, and current overall HRQOL.

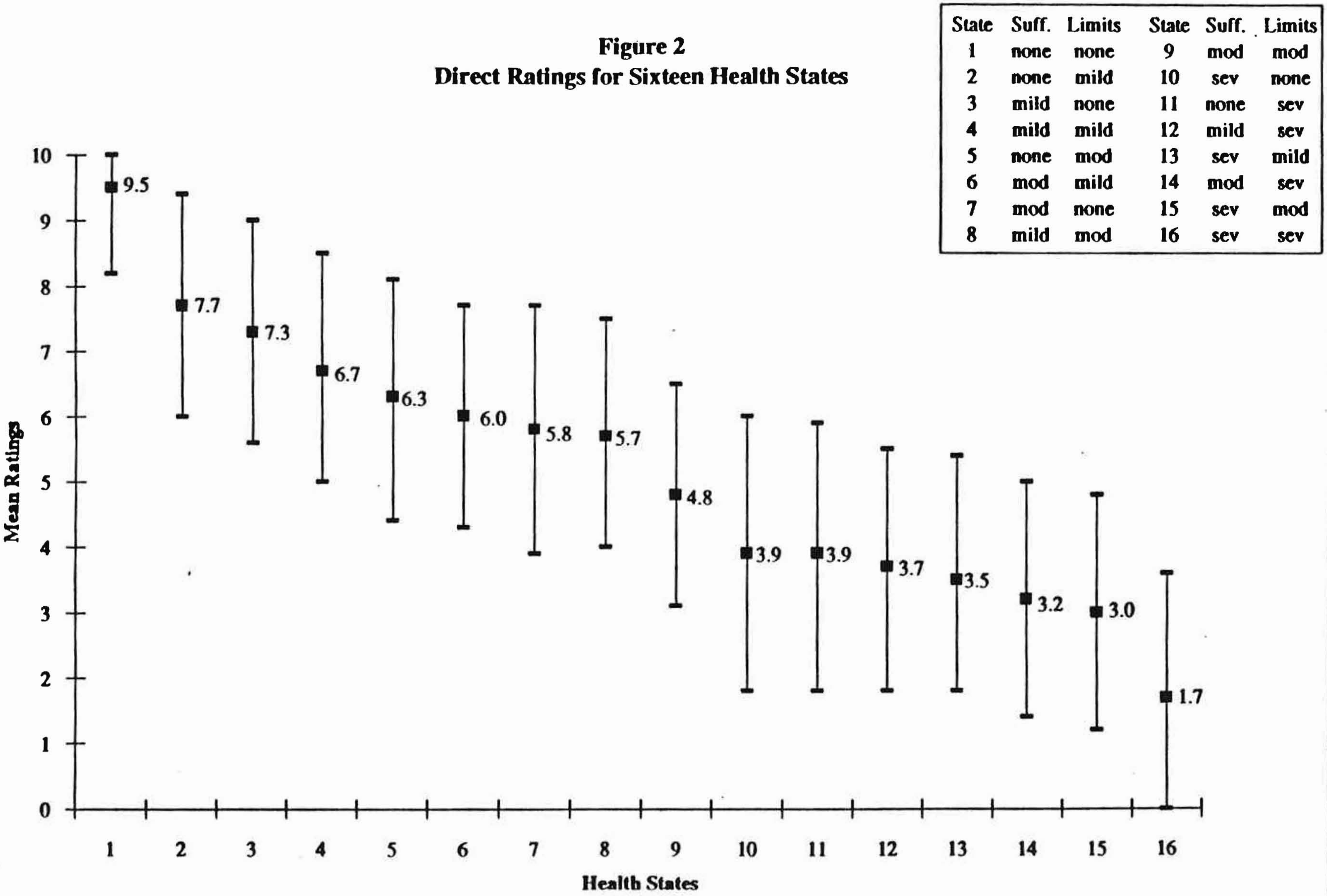
Multivariate statistical analysis showed significant associations between the set of four current HRQOL indices and all three preference rating sets. Association was strongest with the direct ratings. This makes sense, because the comparative nature of items in the other two sets should help reduce the confounding effect of current overall health status.

The strongest association was between the single-item overall current HRQOL rating and the direct rating items; direct ratings for each of the 16 scenarios had significant individual correlations with this variable (average $|r| = .15$). Mean health state ratings of subjects whose current HRQOL was 9-10 (N = 316) was 5.4 (SD = 2.1); ratings for subjects with HRQOL < 8 (N = 144) averaged 4.7 (SD = 2.0). This difference is greater than the one-half rating point which we considered *a priori* to be clinically significant.

Current emotional outlook correlated significantly with ratings on nine of the 16 direct rating scenarios. By comparison, overall current HRQOL correlated significantly with only two of the 13 paired comparison items and three of the 12 time-tradeoff items, and emotional outlook was not significantly correlated with any item of the paired comparison or time-tradeoff sets.

The results demonstrate that subjects' current quality of life and emotional outlook (but not their level of perceived physical suffering or limits on daily activities) may affect their preference ratings for various health states. This finding indicates that emotional outlook and perceived overall quality of life act similarly to "internally calibrate" subjects' ratings of health states. As noted, this effect was most pronounced with use of the direct rating task. In the

Figure 2
Direct Ratings for Sixteen Health States



State	Suff.	Limits	State	Suff.	Limits
1	none	none	9	mod	mod
2	none	mild	10	sev	none
3	mild	none	11	none	sev
4	mild	mild	12	mild	sev
5	none	mod	13	sev	mild
6	mod	mild	14	mod	sev
7	mod	none	15	sev	mod
8	mild	mod	16	sev	sev

analyses reported below, when a current HRQOL item was significantly associated with a demographic variable, the former was treated as a confounding variable in analyses involving the latter.

Demographic Differences

The demographic variables considered were ethnicity, age, and gender. In considering ethnicity we examined only African-American and Caucasian subjects, other groups being insufficiently represented for statistical inference. We decided not to create a "non-Caucasian" category, which we believed would be too heterogeneous for our present purposes. Ethnicity was significantly associated with age and current emotional outlook, and associated at a near-significant level (p = .057) with gender; after controlling for these possible confounds with partial correlations, ethnicity remained significantly associated with five direct-rating items. Caucasians rated these items lower (less desirable) than African-Americans. The mean difference across the five items was .54, (Caucasian mean = 2.91, African-American mean = 3.45). Each of these items includes either severe physical suffering or severely limited activities, or both.

Ethnicity was significantly associated with only two paired comparison items; this association remained significant after controlling for the covariates. The first of these two items compared (no physical suffering + severely limited activities) with (mild physical suffering + moderately limited activities). The second compared (no physical suffering + moderately limited activities) with (mild physical suffering + no activity limitations). For both comparisons, Caucasians showed a greater preference for the second scenario. The results suggest that the Caucasian subjects were more willing to accept the increment from none to mild physical suffering in exchange for fewer activity limitations.

Age Differences

We next looked at whether preferences vary by age. Multivariate analyses revealed significant overall association between respondent age and (i) the direct ratings, (ii) the paired comparison ratings and (iii) the time-tradeoff ratings. Age was significantly associated with ethnicity (African-American vs. Caucasian), sex, and current activity level; the association of age with (i), (ii), and (iii) remained significant after controlling for these possible confounds.

Five direct rating items were significantly correlated with age; four of the correlations remained significant after controlling for the covariates. For each of the four items, older subjects tended to give higher preference ratings. Three of the items involved moderate activity impairment. The results suggest that older subjects are generally less averse to moderate physical suffering and moderate activity impairment. This is not surprising, and may reflect a generally lower comparison level for baseline functioning.

Three paired comparison items were significantly correlated with age; two remained significantly correlated after controlling for the covariates. Older subjects showed greater willingness to accept the change from no to mild physical suffering or mild to moderate physical suffering in return for the change from severe to moderate activity limitations. This may be related to the previous result, i.e., that older subjects rate moderate activity limitation better overall.

One time-tradeoff item significantly correlated with age; the effect was significant controlling for covariates. The item suggests older subjects may be more willing to accept the change from no to mild activity limitation in exchange for a slight increase in life span.

Gender Differences

The final demographic variable evaluated in this study was gender. Using the direct rating task, respondent gender was significantly associated with (i) the direct rating items, (ii) the paired comparison items and (iii) the time-tradeoff items. After controlling for age, ethnicity and current activity limitations, however, only (ii) remained significant; the association of gender with (i), however, was near significant (p = .057).

Despite the significant multivariate association, only two direct rating items had significant associations with gender that remained after adjusting for the covariates. Females rated the condition (no physical suffering + no activity limitations) better (9.7 vs. 9.2) and the condition (severe physical suffering + severe activity limitations) worse (1.5 vs. 1.9) than males. A possible explanation is that males are less likely to give extreme ratings--either high or low.

As before, we observed fewer differences across gender with the paired-comparison task compared to the direct rating task. Indeed, no individual item among the paired-comparison ratings was significantly associated with gender when the effect of the covariates was considered (even without considering the covariates, only one paired-comparison item was significantly associated with gender).

As a final check for systematic effects of demographic factors on ratings, we calculated for each subject an index [alpha] = P / (P + A), where P is the variance accounted for by physical suffering and A is the variance accounted for by activity impairment. The index provides a rough estimate of the importance of physical suffering compared to activity limitation in determining a subject's preference ratings. For the entire population, [alpha] had a mean of .51 and a median of .52, showing that, overall, physical suffering and activity impairment were of near equal importance. The [alpha] index was not significantly associated with age, ethnicity, or gender.

Examination of the paired-comparison data confirmed this result. We calculated the mean of these ratings, reverse-scoring when necessary so that a mean above zero reflects overall preference for less suffering (even at the expense of more activity limitations) and a mean below zero reflects overall preference for fewer limits on activities (even at the expense of more suffering). Mean rating for the 13 paired-comparison scenarios on the 0-10 point scale was 0.12, again confirming near-equivalent importance placed on the two dimensions. Again, no significant differences in mean rating were observed across demographic lines in a multiple linear regression analysis. Note that this rough equality in value between the two dimensions means that the number of items in HRQOL questionnaires should be roughly balanced between these two dimensions.

Effect on Ratings of Experience with Illness

Continuing our search for systematic differences in preference patterns, we next evaluated the effects of having specific medical or disabling conditions on health state preferences.

Specific items in the questionnaire asked about the presence or absence of 17 health conditions, which we later grouped into three broad categories: (1) major medical illnesses, (2) chronic symptomatic conditions, and (3) disabling conditions. The items in the major medical illness category (with endorsement rates in parentheses) were: cancer (1.7%), diabetes (1.2%), heart disease (6.3%), and other major medical problem (12.8%). The chronic symptomatic condition items were: allergies (9.1%), angina (5.8%), arthritis (33.3%), asthma (10.2%), back problems (29.3%), chronic bowel inflammation (4.6%), and ulcer (5.2%). The disabling conditions were: stroke and other neurological disorders (4.7%), incontinence (9.0%), confinement to wheelchair (2.9%), trouble with vision (3.5%), cognitive impairment (20.8%), and other disability or handicap (12.5%).

Multivariate analyses showed no significant association between any of the three broad diagnostic categorizations and either the direct rating, the paired comparison, or the time-tradeoff rating sets. Additional analyses of possible associations between the 17 specific health conditions and the item sets showed several significant associations, but the total number of associations was within that expected by chance alone.

As a final test for preference differences, we identified several subgroups based on demographic and clinical factors, including [number of subjects in each group shown in brackets[]:

(1) age < 40 [206],

(2) age > 65 [171],

(3) African-American [107],

(4) Caucasian [414],

(5) male [228],

(6) female [358],

(7) >50 years old, Caucasian, and female [105],

(8) <50, African-American, and male [21], and

Subjects who reported having in the preceding four weeks:

(9) no illness or conditions [183],

(10) one or more major illnesses only [22],

(11) one or more symptomatic illnesses only [148],

(12) one or more disabling conditions only [37],

(13) one or more conditions in two major diagnostic groups [206],

(14) one or more conditions in all three major diagnostic groups [57],

(15) quality of life of 9 or 10 [315],

(16) quality of life of 6 or less [86],

(17) no physical suffering [212],

(18) moderate or severe suffering [99],

(19) no activity limits [370],

(20) moderate or severe limits on activities [65],

(21) good outlook on life [425]

(22) poor or very poor outlook on life [28].

Relatively few cases were observed in which a subgroup's mean rating was different from the average for the entire sample to a clinically significant extent (>= 0.5 rating points). More importantly, perhaps, no cases were observed in which a subgroup had a significant preference for one state over a second state (either statistically, i.e., *p* < .05 on a paired t-test, or clinically, i.e., 0.5 rating point) when the entire sample significantly preferred the latter state. Nor were any cases found where a subgroup had a significant preference between a pair of states when the entire sample was indifferent between those two states. Several cases were observed, however, in which certain subgroups did not significantly prefer one state over another despite the entire sample having a significant preference between those states. (This phenomenon may be partly explained by the smaller sample sizes within subgroups and resulting lower statistical power). For example, the entire sample and most subgroups significantly preferred the state (no suffering + mild limits) over (mild suffering + no limits). Six subgroups were indifferent between these two states, however: subjects with (1) quality of life rated as 6 or less, (2) moderate or severely limited activities, (3) poor or very poor outlook on life, (4) one or more major illnesses, (5) one or more disabling conditions, (6) one or more illnesses or conditions in all three major diagnostic groups. This example typifies a more general finding: that the subgroups who were indifferent between states that the entire sample had a significant preference between were often more-impaired in terms of conditions, illnesses, or reported quality of life than the average subject.

Figure 3 shows the ratings assigned to two representative health states by several of the patient subgroups listed above.

Preference Structure

The above analysis concludes our search for preference subgroups. In a final set of analyses, we examined the structure of the mean preference ratings derived from both direct rating and paired-comparison tasks. This analysis follows the analyses in our second pilot study.[30]

Figure 4 shows mean ratings on the 4 x 4 = 16 direct rating items across all subjects. As noted above, the scale values derived from the paired-comparison analysis correlated highly with the direct ratings ($|r| = .94$).

Figures 4 has several noteworthy features. First, the bottom line (severe physical suffering) has a steeper slope (reflecting increasingly negative life quality) going from moderate to severe activity limitation than from no to mild or from mild to moderate activity limitations; this effect also occurs in the lines for moderate and mild physical suffering. An analogous effect is shown by the comparatively large gap between the bottom line (severe physical suffering) and the other three lines. For both physical suffering and activity limitation, then, the none, mild, and moderate categories are roughly evenly spaced, with the severe condition more distinct.

Second, the slope of the line denoting moderate physical suffering increases slightly

Figure 3
Subgroup Ratings of Two Health States

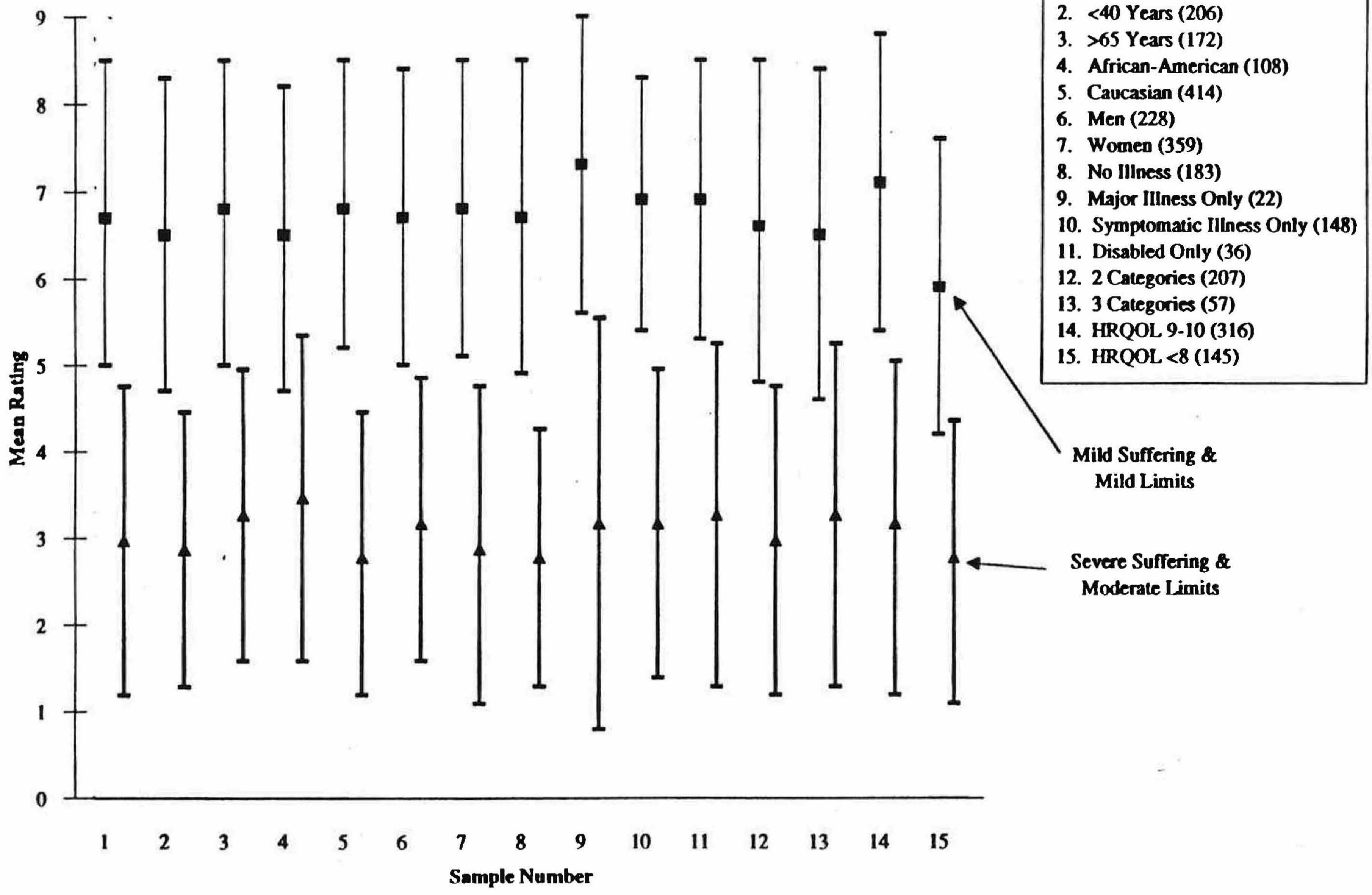
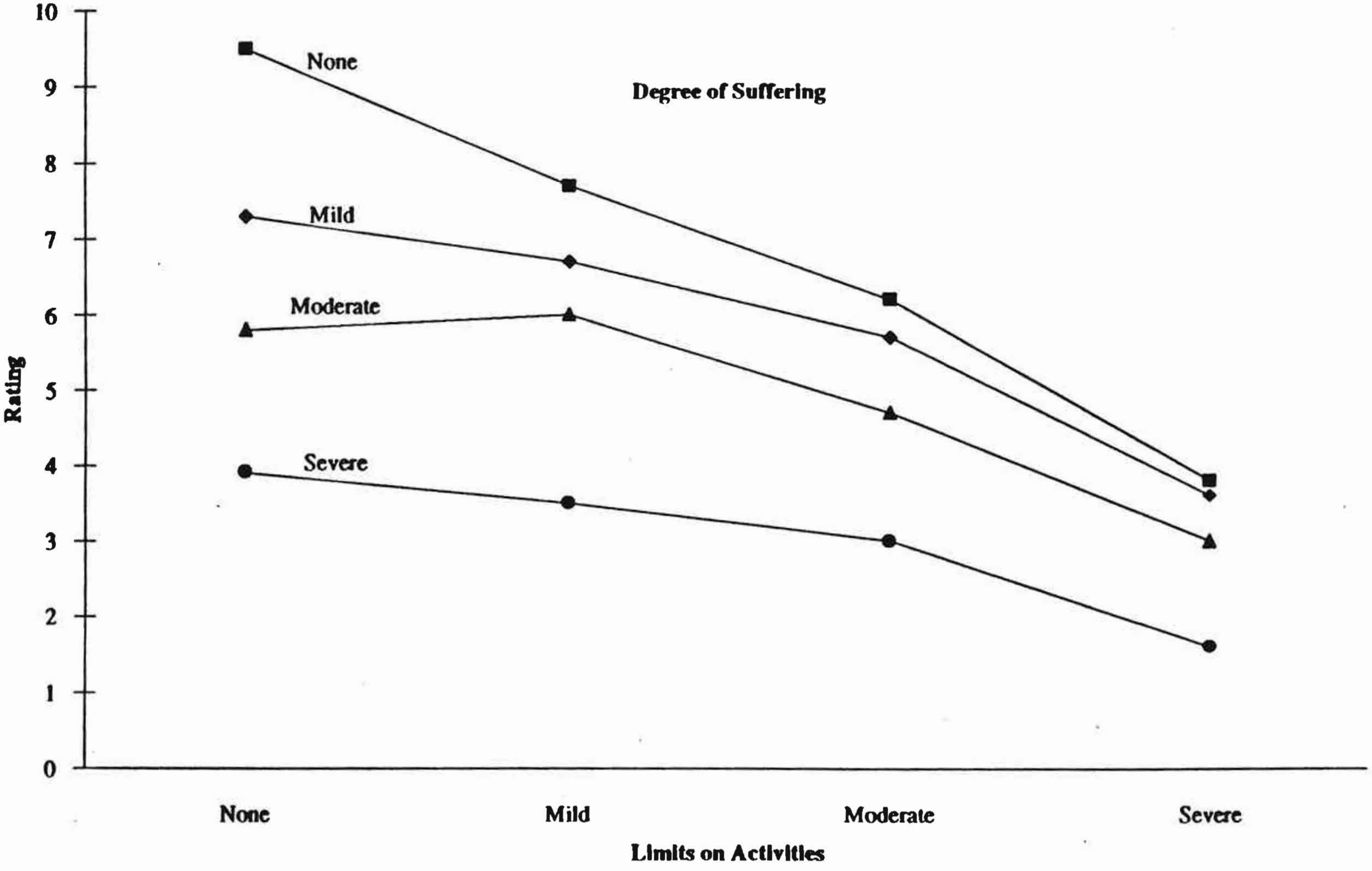


Figure 4
Preference Structure



from the activity impairment level categories of none to mild; this appears to suggests that, given moderate physical suffering, subjects are essentially indifferent to mild versus no activity limitations. This finding is an example of the phenomenon we call "paradoxical indifference."[30] We discuss this finding further below.

A third feature of the figure is the lack of parallelism between the lines. Parallel lines would suggest a simple additive model for integrating negative life quality on the two dimensions--that is, overall negative life quality for a scenario would be a simple sum of the negative life quality associated with the level of suffering and the negative life quality associated with the level of activity limitation. Instead, the lines have a "fan-shaped" structure such that each successive line's slope increases as one goes from bottom (severe suffering) to top (no suffering). This may indicate an asymptotic effect in preference formulation or reporting: if negative life quality is already very poor, additional negative factors may have relatively little effect. Whether this reflects true preferences or is an artifact of the measurement procedures is not clear.

IV. Discussion

The most important finding emerging from our study is the clear absence of systematic differences in comparative preferences for health states across demographic or clinical lines. Very few differences in paired comparison and time-tradeoff ratings were observed. Even with direct ratings, the substantial majority of health states were valued comparably by all subgroups. Furthermore, no significant differences in comparative preference direction emerged when directly rated states were examined in a pair-wise manner. Thus, concerns about systematic preferences differences appear to be unfounded.

Our study has several limitations. First, we did not use random sampling to find our subjects, which raises the possibility of bias and calls into question the generalizability of our results. However, we did make a concerted effort to reach populations not typically captured by preference exercises. Accordingly, we reached a relatively high number of low-income, minority, ill and disabled individuals. Second, our subjects all came from one State, New Jersey, raising similar questions about generalizability. We would be surprised, however, if our subjects were systematically different from similarly situated people in other states. Third, despite enrolling almost 600 subjects many subgroups contain far fewer subjects, thus limiting the statistical power to detect differences in preference patterns.

Preferences and Outcomes Management

The overall question addressed in this discussion paper is whether and how health outcome information can be used to determine "what works" in medicine -- and to design a fair and affordable resource allocation policy. Note that unwarranted attention to the few areas of preference differences could produce highly problematic conclusions. For example, if Caucasians were believed to be more averse to severe suffering and severe activity limits than African-Americans (a conclusion supported by the results of the direct rating task, but not the paired-comparison analysis), health care services that reduced severe suffering or limits in activities might be deemed less desirable for African-Americans. This result is, of course, unacceptable from a social and ethical perspective.

Another troublesome aspect of our findings, discussed briefly above, is the phenomenon of paradoxical indifference. Paradoxical indifference (PI) occurs when two health states are not rated as significantly different (p > .05) despite the fact that one state dominates the other (e.g., mild suffering plus mild limits vs. mild suffering plus moderate limits). When this phenomenon originally appeared in our second pilot study[30] we hoped it would "go away" when larger numbers of subjects were enrolled. This hope was partially realized, in that only one instance of statistically significant PI was noted using the direct rating task (compared with five instances in the pilot study). Specifically, we found that subjects were indifferent (using the direct rating task) between no and mild activity limitations when suffering was moderate. In fact, subjects actually preferred to have mild activity limitations to a statistically significant extent (p = .02). Mean rating for this putatively more-impaired state was 6.00, SD = 1.94, versus 5.83, SD = 1.70 (N = 591) for moderate suffering + no limits.

The observed preference for (or indifference between) mild activity limitations (versus no limits) when suffering is stipulated to be moderate or severe might be explained simply by noting that most people probably would expect or desire to "take it easy for awhile" when they are experiencing significant pain or other form of physical suffering. This finding cannot be explained by the presence of the cartoon figures, because the same preference pattern was observed in our prior study when no figures were used. We discuss at more length in our previous paper[30] the difficulties entailed by paradoxical preferences for resource allocation planning.

Perhaps the most important finding concerning the structure of preferences is the extra weight placed on avoidance of severe suffering or severe limits on daily activities. Thus, services that offer relief of (or improvement in) severe suffering or activity limitations should receive highest priority in systems of health care resource allocation. This result may seem self-evident, but our results lend what might be an important element of empirical support to such a policy.

In Part B of this discussion paper we describe the results of administering our brief four-item questionnaire to a cohort of some 400 cancer patients. The changes over time in HRQOL reported by those patients is scored according to the preferences derived in this part of the study. We model how the use of observed HRQOL outcomes might be used in a system of outcomes management for purposes of resource allocation.

Part B

Questionnaire Validation

I. Introduction

In Part A of this discussion paper we described how a brief, generic questionnaire might be used to measure patients' health-related quality of life (HRQOL) over time on a large-scale, national basis. Such a system of "outcomes management"[1] could be used to ascertain the effectiveness of health care services. This information, in tum, could improve clinical decision-making and help society ensure universal access to treatments and procedures that "work," while curtailing public spending for those that do not.

Determining "what works" depends on public values concerning the outcomes of care.[17] In Part A we reported our experience with calibrating a generic four-item questionnaire -- the Quality of Life and Health Questionnaire (QLHQ) -- according to the preferences of about 600 individuals from various walks of life. We did not detect any systematic differences in preferences based on demographic or clinical factors.

In Part B we describe our experience with the QLHQ in a cohort of 400 cancer patients. The main purpose of the present study was to test the validity of this brief questionnaire under conditions amenable to its use in a system of outcomes management. Therefore, we used a mail-in format designed to maximize response rate, as discussed in Part A. Patients completed the QLHQ at the time of initial enrollment and again three and six months later.

A major problem with all quality of life research is the lack of an objective "gold standard" against which to measure the validity of patient self-reports. In a sense, patients' statements about how they feel about the quality of their own lives could be considered the gold standard itself. After all, can a patient think he has a good quality of life, and be wrong? Can he have a good quality of life without knowing it? Perhaps not, but patients may nonetheless get confused about meanings, may answer in ways they believe will please the interviewer or physician, or may not wish to share their true feelings. In addition, it is well known that patients' evaluations of their health depend as much or more on psychological or emotional factors than on objective indicators of health or function.[45] [46] For these reasons, it is necessary to make some effort to assess the extent to which patients' self-reports reflect their "true" health-related quality of life. Only then can one hope to assess how well questionnaires capture this truth.

We describe our approach to this problem below.

II. METHODS

Subjects

Subjects were 400 patients with newly diagnosed advanced-stage cancer of various types. Subjects were recruited from cancer clinics at the University of Colorado Health Sciences Center and the Veterans Administration Hospital, both in Denver. We selected these patients for two reasons: because (1) we expected that HRQOL for patients with advanced cancer might change over the succeeding several months due to the effects of the cancer and its treatments, and (2) we believed that this cohort of patients, consisting of many elderly, low-income, and quite ill patients, would provide a reasonable test of the feasibility of periodic, patient-completed HRQOL questionnaires.

The study protocol was approved by the University of Colorado Health Sciences Center human subjects committee, and each participating patient gave informed consent. Patients agreeing to participate in the study completed the QLHQ at the time of initial enrollment. Subsequently, project staff mailed additional copies of the QLHQ to all patients three and six months after initial enrollment. Patients were asked to indicate which of the described four levels within each of the three dimensions of HRQOL best described their status over the preceding week. In addition, patients rated their overall quality of life on an 11-point scale, as shown in Figure 1.

We collected a core set of baseline data on each patient, including: (1) age, (2) gender, (3) type of cancer (e.g., lung, breast), (4) stage of cancer, and (5) presence or absence of heart failure, diabetes, and history of stroke. Shortly after initial enrollment the medical record was obtained to determine what types of treatment, if any, were employed against the cancer.

Assessing Questionnaire Validity

The major issue addressed during this part of the study was the validity of the patient-completed QLHQ as an indicator of patients' "true" HRQOL. As noted earlier, there is no "objective gold standard" for measuring patients' "true" HRQOL. We therefore compared the results of the QLHQ with results of a standardized telephone interview conducted by one of two research assistants. This telephone interview was first conducted two to three days after initial enrollment and again immediately after receipt of the patient's completed questionnaire by mail three and six months thereafter. The interviewer did not know the results of the QLHQ at the time of the interview. If necessary, several attempts were made to contact patients via telephone over a period of at least one week, at different times of the day. Patients were considered lost to follow-up if there was no response to a follow-up letter or to a minimum of five telephone calls.

Telephone interviewers completed a standardized form designed to address the three HRQOL dimensions contained in the QLHQ. Figure 5 lists the major questions asked by the interviewers, who selected the level most closely matching the patients' responses. Detailed follow-up questions were asked as needed to clarify patients' responses. The first fifteen patient interviews were conducted by one interviewer (alternating as to which one), with the other interviewer listening over a telephone extension. (This was done with the patients' knowledge and consent.) Each interviewer completed a response form independently; responses were compared after the interview was completed, and any differences discussed. Responses were nearly identical after the first ten patients, with no more than one answer differing (by one level) on the entire response form. We believe that the ratings of these interviewers, who together conducted over 1,000 standardized telephone interviews, comes as close to a true "gold standard" of patient HRQOL as is likely to be identified.

Interviewers' responses to the questions in Figure 5 were transformed into summary scores reflecting patients' status along each of the three basic dimensions of HRQOL. Within each dimension, each "a" level answer was assigned zero points, each "b" level answer was assigned one point, each "c" level answer was assigned two points, and each "d" level answer was assigned three points. The sum of the total number of points scored within each dimension was used as an overall index for that dimension. The degree of correlation between each summary score and the corresponding patient response on the QLHQ was measured using Pearson's product-moment correlations. Overall HRQOL scores obtained from patients and interviewers were also compared in this way.

Correlations among measures were evaluated using the multitrait-multimethod (MTMM) analysis described by Campbell and Fiske.[30 47] In this setting, the "traits" to be measured were patients' physical suffering, limits on daily activities, and emotional outlook. Each of these three traits was measured at the time of initial enrollment and three and six months later, producing a total of nine separate and distinguishable "sub-traits." Thus, for purposes of this analysis we treat physical suffering at three months as a trait distinct from physical suffering at six months. This construction is based on the assumption that a valid questionnaire should be able to detect change in physical suffering (and in other dimensions of HRQOL) over time. This characteristic has been termed "responsiveness,"[48] but probably is best considered simply as "validity-over-time."[49]

Each of the nine traits was measured using two methods: the original QLHQ patient self-reports and the aggregate trait scores obtained through the follow-up telephone interviews, as

FIGURE 5

STANDARDIZED TELEPHONE INTERVIEW

Physical Suffering

1. How much pain have you experienced in the last week?

   a) None
   b) A little
   c) Quite a bit
   d) Severe pain

2. How much nausea/vomiting have you had in the last week?

   a) None
   b) A little
   c) Quite a bit
   d) Severe nausea/vomiting

3. Does pain or other symptoms keep you awake at night?

   a) Never or rarely
   b) Sometimes
   c) Often
   d) Very often or almost every night

Daily Activities

1. How often are you able to do things you enjoy or that are important to you?

   a) Very often or almost every day
   b) Often
   c) Sometimes
   d) Never or rarely

2. How much help do you need getting around?

   a) No help
   b) A little help
   c) Moderate amount of help
   d) A lot of help

3. How much assistance do you need taking care of daily needs such as getting dressed or eating?

   a) No help
   b) A little help
   c) Moderate amount of help
   d) A lot of help

FIGURE 5 (CONT.)

STANDARDIZED TELEPHONE INTERVIEW

Emotions and Outlook on Life

1. How often this week have you felt depressed and upset?

a) None
b) A little
c) Quite a bit
d) Very often/most of the time

2. How much of a problem have you had this past week with sleeping at night because of nervousness or anxiety?

a) None
b) A little
c) Quite a bit
d) Very often/major problem

3. How happy are you feeling these days?

a) Very
b) Somewhat
c) A little
d) Not at all

Interviewer's global estimate of patient's quality of life (0-10 scale, 0-worst possible quality of life, 10-best possible quality of life).

0......1......2......3......4......5......6......7......8......9......10

(Circle one number)

described above. In this test of questionnaire validity, correlations between two methods' measures of a given trait (i.e., monotrait, heteromethod values) should be higher than (1) the correlations between these two methods' measures of separate traits (heterotrait, heteromethod) and (2) a single method's measure of different traits (heterotrait, monomethod).

We also conducted a separate MTMM analysis of the validity of the global 0-10 HRQOL item. Here, the traits to be measured were overall HRQOL at zero, three, and six months. The different methods of measurement were (1) patients' self-reported HRQOL on the 0-10 item, (2) the sum of patients' scores on physical suffering, limits on activities, and emotional outlook, (3) the interviewers' 0-10 estimate of the patient's HRQOL, and (4) the overall sum of the interviewer's aggregate index scores on suffering, limits, and outlook. We hypothesized that the correlation coefficients among measures taken at a single time point should be higher than correlations among measures taken at different times.

MTMM analysis is generally considered to provide evidence of construct validity, i.e., the index measure should correlate more highly with measures hypothesized to assess the same construct than with measures hypothesized to assess different constructs. In this case, however, because the second measurement method was the follow-up telephone interview -- which we took as an indication of "true" HRQOL -- we believe that the MTMM analyses can also reasonably be considered to reflect concurrent or criterion validity.

Our second approach to assessing the validity of the self-reported quality of life measures was to compare the results of these measures with subsequent observed mortality rates. It is well-established that global measures of health are highly correlated with mortality risk.[50] [51] [52] [53] [54] For example, Mossey and Shapiro[54] found that people whose self-rated health was poor experienced two-year mortality rates about three times higher than those who rated their health as excellent. We compared patients' baseline status on each HRQOL dimension with mortality rates observed six months after enrollment. We hypothesized that patients reporting poorer initial HRQOL would experience higher mortality rates.

To assess the degree of association between baseline HRQOL and mortality, we calculated chi-square values on n x 2 contingency tables on mortality and each HRQOL dimension (including overall quality of life). Significance levels were set at $p < .05$. We then compared the mortality rates of patients who reported no problem versus some degree of problem with (1) suffering, (2) limits on activities, and (3) outlook at baseline -- as well as patients above versus below various cutoff points in initial overall quality of life and health state weights. To obtain this latter parameter we placed patients into one of 16 health states (ranging from "no suffering and no limits" to "severe suffering and severely limited") by combining their responses to the physical suffering and daily activity items, as described in Part A of this discussion paper. We assigned weights to these states based on the results of the direct-rating task described in Part A. Differences in mortality rates between the various pairs of comparison groups were assessed using simple chi-square analyses, with significance set at $p < 0.05$. Finally, we conducted multivariate analyses (i.e., logistic regression) to assess the effect of baseline HRQOL dimensions on mortality after controlling for other significant predictor variables, including demographic, clinical, and treatment variables. Significance was again set at $p < .05$ in these analyses.

Our final set of analyses assessed the extent to which reported changes in HRQOL over time (i.e., quality-of-life outcomes) could be accounted for (or predicted) by baseline factors, including demographic and clinical characteristics and patients' self-reported HRQOL. We examined the effect of age, gender, hospital (VA vs. University), stage of cancer, treatment selected, and self-reported HRQOL on changes in quality of life over time. For these analyses we created two sets of HRQOL-outcome variables. First, we created change scores for each dimension of HRQOL by subtracting the baseline rating on a given dimension (including overall quality of life) from the rating obtained on that dimension six and twelve months later. Second, we created a separate set of change scores based on the difference in preference weights of the health states reported at baseline and at six and twelve months.

Change scores on individual HRQOL ratings and on health-state weights were treated as continuous outcome variables. We first assessed the univariate association of these scores and each candidate predictor variable (i.e., demographic and clinical characteristics, treatments, and baseline HRQOL ratings) using a simple chi-square test; variables found to be significantly associated with change scores were entered into multiple linear regressions in order to determine the extent to which they accounted for observed changes in quality of life after controlling for other significant predictor variables. Because of the large number of statistical tests performed in these analyses, we considered an association significant if its p-value was less than or equal to 0.01.

III. Results

Figure 6 summarizes the characteristics of our patient sample. Six patients died or dropped out prior to the first telephone interview, leaving 394 patients in our sample.

Questionnaire Validity

We first examined the extent to which patients' self-reported HRQOL ratings correlated with ratings obtained on the corresponding dimensions during the subsequent telephone

FIGURE 6

CHARACTERISTICS OF PATIENT SAMPLE

N = 394		Type of cancer	
Female	66%	Lung	14%
Age (Median = 60)		Breast	10%
18-35	9%	Gastrointestinal	11%
36-50	22%	Genitourinary	23%
51-65	35%	Lymphoma	10%
66-80	31%	Leukemia	6%
Over 80	3%	Other	25%
Diabetes	6%	Stage II	6%
Heart failure	10%	Stage III	47%
History of stroke	2%	Stage IV	47%
Treatment received:			
Chemotherapy	47%		
Surgery	20%		
Radiation	7%		
Hormones	5%		
Immunotherapy	3%		
None	21%		

Note: Some patients received more than one treatment modality. Cancer was classified according to standard staging protocols, with Stage III generally involving regional spread (including lymphatic system) and Stage IV associated with distant metastases.

interviews. As described above, we used the results of these interviews to create summary scores for each dimension of HRQOL. These scores were compared to patients' responses using a multitrait-multimethod analysis. Figure 7 is a correlation matrix among these various measures. Underlined values show the validity diagonals, which represent the convergences of the two separate measurement methods on each of the nine traits (e.g., activity limits at three months). (These are the "monotrait, heteromethod" correlations.) Mean correlations on the validity diagonals was 52.9 (SD 12.7); off-diagonal correlations averaged 32.7. This difference was highly significant (*p* = .001 on a Student's t-test). Also, note that the on-diagonal values were usually the highest value in their respective rows and columns. This is the expected pattern for valid measures. Thus, we conclude that patients' responses to the separate HRQOL dimensions contained in the QLHQ were validated by the separate telephone interviews. That is, patients response to, for example, the suffering item "really did" reflect suffering, rather than emotional status, activities, or some other construct.

Figure 8 shows the results of a separate MTMM analysis on the global HRQOL measure. Again, underlined values represent monotrait, heteromethod convergence on validity diagonals. On-diagonal correlation coefficients averaged 64.2 (SD 10.0); mean off-diagonal correlation was 41.3 (SD 9.5). This difference was also statistically significant (*p* = .0001 on a Student's t-test). Thus, the global 0-10 HRQOL item manifested evidence of excellent validity over time. Again, on-diagonal values were usually the highest value in their respective rows and columns.

Mortality Analyses

We next examined the extent to which initial self-reported HRQOL scores corresponded with mortality risk. By six months after enrollment, 21 of our original 394 patients were lost to follow-up. We believe that most of these patients probably died, but we were unable to confirm this suspicion. Of the remaining 374 patients, 59 (16%) were confirmed to have died within six months. If all 21 patients lost to follow-up are assumed to have died, the overall six-month mortality rate would have been 20%. As a sensitivity analysis of sorts, we used both the proportion of patients confirmed dead at six months and the total proportion of patients who were either confirmed dead or lost to follow-up at six months in our mortality analyses.

As expected, baseline patient self-reported HRQOL was significantly associated with six-month mortality risk. Chi-square values for each of the three 4 x 2 contingency tables created using the three basic dimensions of HRQOL (i.e., physical suffering, limits on activities, and

FIGURE 7

CORRELATION MATRIX OF PATIENTS' AND INTERVIEWERS' RESPONSES TO DIFFERENT HRQOL DIMENSIONS

```
(obs=273)

         |   suff0   activ0    emot0    suff3   activ3    emot3    suff6
---------+---------------------------------------------------------------
   suff0 |  1.0000
  activ0 |  0.5240   1.0000
   emot0 |  0.4009   0.3659   1.0000
   suff3 |  0.2745   0.2409   0.1725   1.0000
  activ3 |  0.2660   0.3245   0.2764   0.4606   1.0000
   emot3 |  0.2080   0.2448   0.3910   0.3939   0.5591   1.0000
   suff6 |  0.3408   0.2567   0.2860   0.4498   0.3723   0.3826   1.0000
  activ6 |  0.2788   0.2808   0.2888   0.3350   0.4348   0.3232   0.6402
   emot6 |  0.1589   0.2252   0.3216   0.2602   0.3191   0.5022   0.5941
psufsum0 |  0.4062   0.3267   0.2414   0.2798   0.2629   0.2765   0.2897
pactsum0 |  0.2592   0.2859   0.2148   0.1339   0.2573   0.2006   0.2211
pemosum0 |  0.2034   0.2135   0.4225   0.2432   0.3131   0.3822   0.2877
psufsum3 |  0.2912   0.2879   0.3210   0.5189   0.4124   0.3961   0.4257
pactsum3 |  0.2391   0.3572   0.2404   0.4199   0.5967   0.4007   0.3660
pemosum3 |  0.2664   0.2483   0.4030   0.3101   0.4694   0.6127   0.3568
psufsum6 |  0.2467   0.1570   0.2752   0.3967   0.3327   0.3558   0.6290
pactsum6 |  0.1841   0.1522   0.1812   0.3337   0.3633   0.2610   0.5187
pemosum6 |  0.1807   0.1681   0.3317   0.2787   0.3340   0.4487   0.5814

         |  activ6    emot6 psufsum0 pactsum0 pemosum0 psufsum3 pactsum3
---------+---------------------------------------------------------------
  activ6 |  1.0000
   emot6 |  0.5183   1.0000
psufsum0 |  0.2327   0.1837   1.0000
pactsum0 |  0.3007   0.1970   0.3037   1.0000
pemosum0 |  0.2487   0.3071   0.4247   0.3089   1.0000
psufsum3 |  0.3477   0.3015   0.3605   0.1596   0.3043   1.0000
pactsum3 |  0.5029   0.3114   0.2758   0.3422   0.2798   0.3687   1.0000
pemosum3 |  0.3001   0.3916   0.3097   0.1508   0.4688   0.4797   0.4231
psufsum6 |  0.4523   0.4177   0.3875   0.1395   0.2769   0.5360   0.2928
pactsum6 |  0.6281   0.4157   0.2324   0.2276   0.1951   0.3170   0.5309
pemosum6 |  0.5059   0.6481   0.2431   0.1073   0.2958   0.3450   0.2867

         | pemosum3 psufsum6 pactsum6 pemosum6
---------+------------------------------------
pemosum3 |  1.0000
psufsum6 |  0.3824   1.0000
pactsum6 |  0.2587   0.5075   1.0000
pemosum6 |  0.5098   0.5411   0.4932   1.0000
```

Legend.

From QLHQ:
suffx = patient's reported degree of physical suffering at time x
emotx = patient's reported emotional outlook at time x
activx = patient's reported degree of activity limitation at time *x*
qolx = patient's reported overall HRQOL at time *x*

From telephone interview:
psufsumx = interviewer's summary score on suffering at time *x*
pemosumx = interviewer's summary score on emotional outlook at time *x*
pactsumx = *interviewer's summary score on activity limitation at time* x

Underlined values represent monotrait, heteromethod convergences (see text).

FIGURE 8

CORRELATION MATRIX FOR OVERALL HRQOL MEASURES

```
(obs=267)
         |     qol0      qol3      qol6   totsum0   totsum3   totsum6     pqol0
---------+----------------------------------------------------------------------
     qol0|   1.0000
     qol3|   0.4709    1.0000
     qol6|   0.3741    0.5663    1.0000
  totsum0|   0.6618    0.4162    0.2435    1.0000
  totsum3|   0.2598    0.7000    0.4547    0.4021    1.0000
  totsum6|   0.3552    0.4502    0.7156    0.3866    0.5284    1.0000
    pqol0|   0.4753    0.3746    0.2796    0.4892    0.3223    0.3289    1.0000
    pqol3|   0.3385    0.6473    0.5116    0.3908    0.6292    0.5269    0.4691
    pqol6|   0.3450    0.5064    0.7128    0.2882    0.4609    0.6782    0.3432
 ptotsum0|   0.4124    0.4228    0.3120    0.4828    0.4285    0.3781    0.6857
 ptotsum3|   0.3424    0.6654    0.4611    0.4616    0.7234    0.5614    0.3762
 ptotsum6|   0.2826    0.4464    0.6303    0.2950    0.4961    0.7595    0.2863

         |    pqol3     pqol6  ptotsum0  ptotsum3  ptotsum6
---------+--------------------------------------------------
    pqol3|   1.0000
    pqol6|   0.6168    1.0000
 ptotsum0|   0.4300    0.2966    1.0000
 ptotsum3|   0.7244    0.5112    0.5008    1.0000
 ptotsum6|   0.5388    0.7624    0.3767    0.5944    1.0000
```

Legend.

From QLHQ:
qol*x* = Patient's reported overall HRQOL at time *x*
totsum*x* = Sum of patient's scores on suffering, limits, and emotions at time *x*

From telephone interviews:
pqol*x* = interviewer's estimate of patient's overall HRQOL at time *x*
ptotsum*x* = Sum of interviewer's scores on suffering, limits, and emotions at time *x*

Underlined values represent monotrait, heteromethod convergences (see text).

outlook on life) versus mortality were significant at p < .001. Chi-square for the 11 x 2 contingency table on overall baseline HRQOL versus mortality was significant at p = .004.

Figure 9 summarizes the mortality rates of different patient subgroups. In general, patients with higher baseline HRQOL, higher-weighted health states, or lesser degrees of impairment on the three basic HRQOL dimensions experienced six-month mortality rates of about 10% and combined dead/lost-to-follow-up rates of about 14%. Patients with lower baseline HRQOL, lower-weighted health states, or more impairment on the three basic HRQOL dimensions had mortality rates of about 18% and combined dead/lost-to-follow-up rates of about 24%. As shown in Figure 6, most of these differences were statistically significant.

The observed significant effects on mortality rate of baseline HRQOL parameters persisted in multivariate analyses; the only other variables found to be significantly associated with six-month mortality were stage of cancer and presence of lung cancer. Treatment variables did not predict mortality.

Changes in HRQOL Over Time

We next examined trends in patient-reported HRQOL over time. Figure 10 depicts the pattern of results observed for the HRQOL measures, including (1) pain and physical suffering, (2) emotions/outlook on life, (3) limitations in daily activities, (4) overall quality of life, (5) health state weights (i.e., weighted combination of suffering and limits, as described above), and (6) interviewer's global estimate of patients' HRQOL. Scores were transformed to parallel the overall quality-of-life measures, i.e., a value of 10 indicated the most desirable situation (e.g., no suffering or limits on activities) and a value of 0 the least desirable situation (e.g., severe suffering or limits).
It can be seen that the measures changed relatively little over the six-month period.

For our analyses of changes in quality of life we used the change-score derived by subtracting six-month overall quality of life ratings or six-month health-state weights from corresponding ratings and weights at baseline. Figure 11 depicts the distribution of six-month quality-of-life change scores. Using multiple linear regression analysis we determined that the most powerful predictor of change in quality of life or weighted health states was initial quality of life and health state, respectively. Interestingly, higher baseline ratings (meaning better initial HRQOL) were significantly associated with *decreases* in quality of life over time. For example, the 146 patients who were known to be alive at six months and who rated their initial

FIGURE 9

Differences in Mortality and Lost-to-Follow-up Rates

Values on Baseline Parameters	Number Dead	Number Dead or Lost to Follow-up
Overall HRQOL		
>=8	17/163 (.10)[a]	23/169 (.14)[b]
<8	41/210 (.20)	57/225 (.25)
Extent of Suffering		
None	10/105 (.10)[c]	13/108 (.12)[d]
Mild, Moderate, Severe	48/268 (.18)	67/286 (.23)
Extent of Limits on Activities		
None	13/141 (.09)[d]	20/148 (.14)[d]
Mild, Moderate, Severe	45/232 (.19)	60/246
Extent of Problems with Outlook on Life		
None	15/137 (.11)[e]	19/141 (.14)[d]
Mild, Moderate, Severe	43/235 (.18)	61/252 (.24)
Health State Weight		
>=7	14/156 (.09)[f]	22/163 (.13)[f]
<7	44/217 (.20)	58/231 (.25)

Differences significant at:

[a]p=.02 [b]p=.004 [c]p=.04 [d]p=.01 [e]p=.06 [f]p=.005

overall quality of life as "8" or higher, experienced an average decrease in quality of life of 0.9 rating units. By contrast, the 168 patients alive at six months who rated their initial quality of life less than "8" experienced an average *improvement* in quality of life of 0.7 rating points. This difference in change scores was significant at the $p = .0001$ level. Similar patterns were observed for changes in weighted health states.

The inverse association of initial HRQOL and HRQOL outcome persisted in multivariate analyses. Gender was the only other variable found to have a significant effect on overall quality-of-life outcomes. Men experienced an average decrease in overall quality of life of 0.14 rating points, whereas women experienced an average improvement of 0.57 rating points. This difference was significant at $p =.001$. With regard to six-month changes in health-state weights, both gender and stage of cancer were significantly associated with changes in weights in multivariate analyses (in addition to baseline health-state weight). Men experienced an average improvement in health-state weight at six months of 0.10; women improved by 0.96. This difference was significant at $p < .001$.

IV. Discussion

We observed high correlations between self-reported HRQOL on the Quality of Life and Health Questionnaire (QLHQ) and subsequent standardized, in-depth telephone interviews. Results of the MTMM analyses indicate that both the separate dimensions of HRQOL (suffering, activity limits and emotions) and global HRQOL measure provide reasonable reflections of the actual underlying constructs, both at a single time point and over time. In addition, we observed the expected pattern between initial self-reported HRQOL and subsequent mortality. These findings demonstrate that the QLHQ, when completed by patients in a mail survey format, is a valid measure of patients' actual HRQOL.

We achieved high compliance rates in our study; in fact, all but two patients who did not die and who were not lost to follow-up completed all three QLHQ questionnaires at baseline and at three and six months. Although several patients required reminder telephone calls, in general we found that patients were extremely enthusiastic about providing us with information about their HRQOL. Indeed, patients would often "pour their hearts out" to us about their experiences. Although this high level of compliance would not be expected in the setting of large-scale outcome management programs, we beleve that the brief format and easily understood questions would facilitate high compliance rates even in this latter setting.

The finding of significant inverse associations between baseline HRQOL and quality-of-life outcome is probably due in large part to regression to the mean. A related "ceiling" effect on HRQOL is also at work; patients who rate their initial HRQOL very high have little or no room for improvement, but lots of room for worsening. Patients who rate their initial HRQOL lower may have as much or more room to improve than to deteriorate. The significance of this inverse association for purposes of outcomes management is uncertain, although clearly initial quality-of-life ratings or weights must be entered into multivariate analyses of outcomes to control for this phenomenon.

In our study we did not detect any treatment effects on either mortality or HRQOL outcomes after correction for gender and type and stage of cancer. This is not surprising, because our sample contained relatively few patients with a given type or stage of cancer (see Figure 6), and even fewer patients with a given type of cancer who received a particular treatment. Much larger studies, such as are envisioned by Ellwood in systems of outcomes management, would potentially contain thousands of patients with a given type and stage of cancer. With such large numbers of patients, we would often expect to find significant differences in health outcomes across alternate courses of treatment for patients with a given clinical condition. Indeed, such findings form the very basis for the concept of outcomes measurement and management.

In our analyses we used change scores to depict the difference between initial and subsequent HRQOL. Alternatively, the subsequent scores themselves (e.g., six-month overall HRQOL) could be used as the dependent variables in multivariate outcome analyses. Inclusion of initial HRQOL scores in either case should help to minimize the biasing effect of initial HRQOL on outcome inferences.

Use of a Global HRQOL Measure

For reasons summarized above, we believe that the QLHQ can provide a useful and valid basis for regularly surveying large numbers of patients concerning their HRQOL. The generic states depicted in the QLHQ adequately represent, we believe, the range of symptoms and limitations produced by medical conditions and diseases -- as well as the improvements (or worsenings) of symptoms and limitations produced by medical and surgical treatments and procedures. Indeed, when very large samples are obtained, outcome analyses might reasonably be focused on changes in the single, global HRQOL measure contained in the QLHQ. Use of single measure would almost certainly improve compliance rates over even a four-item questionnaire.

Use of a global, generic HRQOL item for purposes of outcomes management raises several questions. First, what about the problem of questionnaire calibration discussed in Part

A of this article? In a sense, however, single-item global measures might be considered "internally calibrated," especially when they are anchored, as ours was, by terms such as "best possible quality of life" and "worst possible quality of life." For example, a patient who selected a 7 as most closely representing his current HRQOL would be saying, in effect, that he considers his HRQOL about 70% as desirable as the best possible quality of life. This element of desirability closely parallels the notion of preferences and values.

Beyond this conceptual similarity between global measures and preference-weighted health states, we observed in this study empirical support for the comparability of these two types of measures. Specifically, our global HRQOL measure closely paralleled the health state weights produced by assigning the empirically derived scale values (per Part A of this article) to the various combinations of physical suffering and activity limits reported by patients. This similarity was apparent both in the multivariate and mortality analyses reported in the Results section and in the observed changes in patients' HRQOL over time, as depicted in Figure 10. Global patient-reported HRQOL fell in the middle of the seven measures depicted in that Figure.

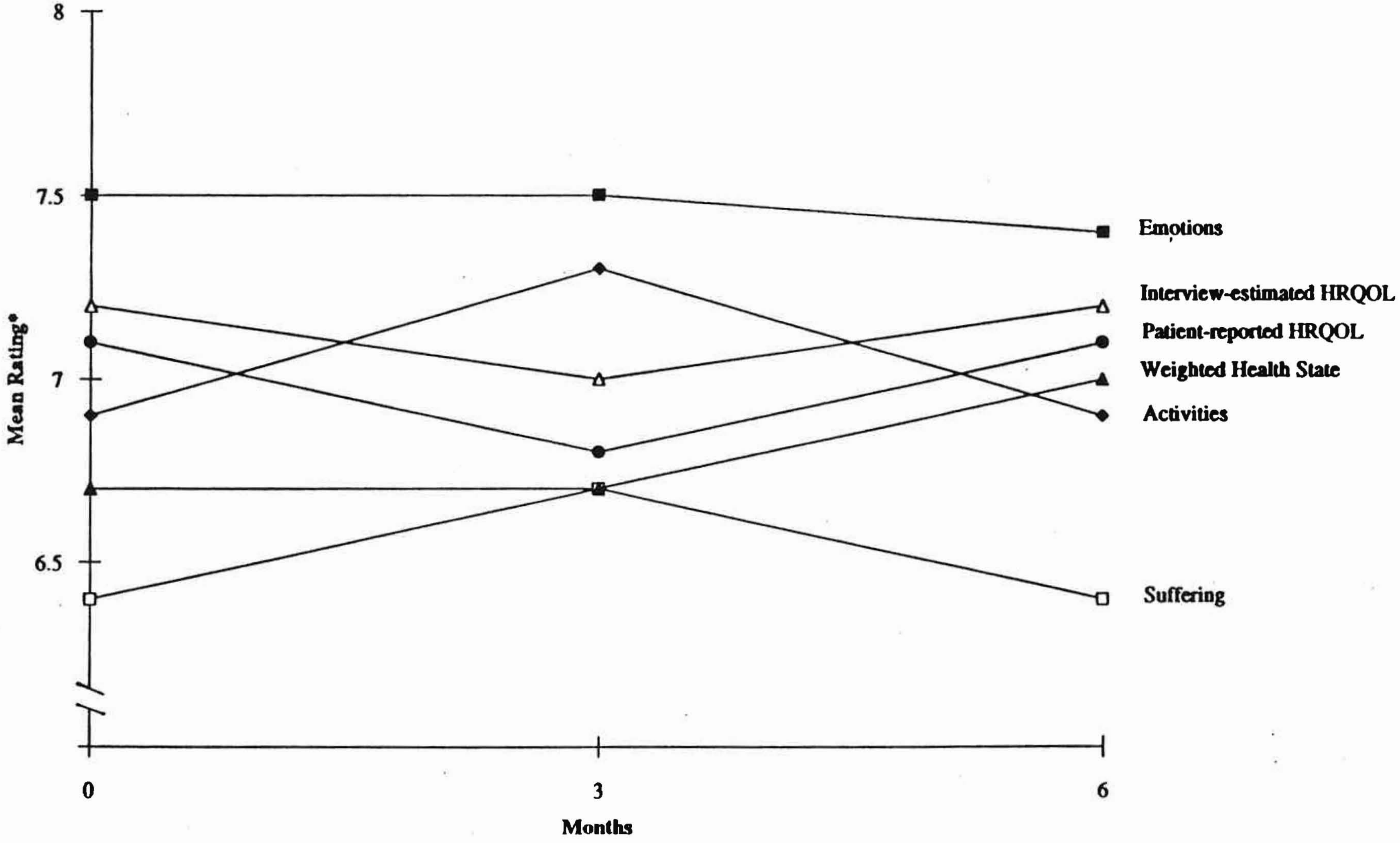
Other investigators have advocated the use of single-item measures for capturing patients' HRQOL over time, including Gough et al., who recommended that patients be asked a single question: "Overall, how would you rate your quality of life today?"[55] These investigators note that such global indicators correlate well with multi-dimensional constructs and that more complex conceptualizations of quality of life require an incre

ase in the complexity of patient outcome questionnaires. Moreover, many multi-item questionnaires are validated by the extent to which the aggregate score on all items correlates with a single "overall quality of life" item. Why not, then, use the single item in the first place?

An additional argument in favor of single-item questionnaires was mentioned above, namely the fact that responses to single measures correlate significantly with mortality risk.[50-54] We observed similar correlations in the present study. This universal finding is further evidence that single, global measures can provide valid depictions of patients' overall state of health.

Alvan Feinstein identified additional advantages of single-item questionnaires, noting that longer, multiple-item questionnaires "are seldom satisfactory for individual patients, because the main focus of clinical concern or management may be lost, obscured, or not included among all the other information that surrounds it."[56] A single-item, "global index," on the other hand, may be the best way of letting the patient decide and indicate the rating for

Figure 10
Behavior Over Time of HRQOL Parameters



* Transformed where necessary so that 0 = worst situation and 10 = best situation (see text).

whatever he or she chooses to emphasize as the important features of life. Furthermore, if the indexes have an adequate number of categories, they may be particularly desirable for letting patients use their own criteria to determine changes in their status.[56,(p.MS53)]

Finally, good to excellent reliability and validity have been observed for several single-item HRQOL questionnaires,[57] [58] including the Delighted-Terrible Scale,[58] [59] the Faces Scale[58], and the Ladder Scale.[60]

Induction Problems

Even when data are aggregated into massive data sets, complete with global HRQOL measures, the results of retrospective outcome analyses may be difficult to accept at face value. It is widely appreciated that all non-experimental (i.e., non-randomized) study designs are vulnerable to a variety of biases, particularly with respect to the presence of unmeasured confounders.[61] For this reason, the medical records used to supply data to systems of outcomes managment should contain as much information as possible about comorbidities, severity of illness, and other variables related to outcomes in order to permit statistical adjustment of observed outcomes and to strengthen causal inferences.[62]

Even under the best of circumstances, however, it will be difficult to forge what Ellwood called the "link [between] the actions and observations of thousands of health professionals with [the outcomes of] millions of patients."[1] Ellwood acknowledges that "major questions surround attempts to measure the impact of medical care on quality of life," including the potential for bias, but he argues that outcomes management "would generate information about the results of the natural, seemingly random variations in practice style." This overstates the methodological case for what would be, in effect, a massive case-control study. In Part C we discuss further the problem of interpreting observational outcome data.

In summary, we have demonstrated that very brief, mail questionnaires, perhaps consisting of a single, global measure of HRQOL, can serve as a valid basis for monitoring the HRQOL outcomes associated with diseases, conditions, treatments, and procedures. We also determined that additional items concerning physical suffering, functional status, and emotional outlook can obtain valid information concerning these respective dimensions. Whether the additional information gained is worth the potential decrease in compliance is unknown at this time.

We echo Ellwood's call for the routine collection of HRQOL data and urge policy-makers to make creation of large data bases a high priority.

Part C

Problems with Interpretation

I. Introduction

In Parts A and B of this discussion we reported that a brief mail survey, consisting of only four questions and calibrated according to empirically derived preferences, was able to provide valid information about the health-related quality-of-life (HRQOL) outcomes of patients with a serious medical condition. Information of this type, collected regularly and systematically, could become the basis for a system of "outcomes management,"[1] in which efforts were made to determine "what works" in medicine. Such a system would provide an efficient and valuable alternative (or supplement) to randomized controlled trials and to more traditional types of non-experimental studies. In this final Part, we discuss some of the philosophical and practical issues that would arise during creation and use of a system of outcomes management.

In Part B we concluded that a single, global HRQOL item would probably suffice for purposes of measuring outcomes in a large system of outcomes management. We observed excellent construct and criterion validity of such a global item in a cohort of cancer patients, with "objective" interviewers' careful assessments of patients' HRQOL serving as the "gold standard" of validity. Our 0-10 global HRQOL measure tracked very closely with patients' responses to items on physical suffering, activity limits, and emotional outlook -- and with empirically derived preference weights for combinations of problems with suffering and activity limits (as described in Part A). These separate HRQOL items also manifested excellent construct validity. Finally, we found that determinants and behavior of the overall HRQOL measure were very similar to that of the preference-weighted health states reported by patients.

In constructing the ideal survey instrument for large systems of outcomes management, one could use a single, global item or include a few additional items, as we did. Responses to all items could be "blended" together to form an overall estimate of HRQOL. Although this approach *might* increase reliability of the HRQOL estimate somewhat, we saw no evidence that the accuracy, or validity, of this estimate provided by the global item would be enhanced by the presence of additional items. Moreover, increasing the number of items would likely result in lower response rates. The optimal number of items remains an open question for now, but it seems clear in any case that a few items would be sufficient. This is fortunate, in discussed in previous Parts and again below.

Even granting that single, global items can produce valid measurements of overall HRQOL, it still may seem difficult to believe that such measures could provide the basis for conclusions about the effectiveness of particular procedures and treatments. In this Part we discuss two major issues with respect to the problem of interpreting change in global HRQOL outcome: (1) how to strengthen causal inferences drawn from observational HRQOL data, and (2) how to identify, using statistical procedures performed using global HRQOL data, different "types of patients" who either do or do not receive significant benefit from some treatment or procedure. We conclude that global HRQOL data, if properly obtained and interpreted, can indeed provide a legitimate basis for large-scale outcome studies of the type envisioned by Ellwood some five years ago.

II Strengthening Causal Inferences

In Part A of this discussion paper we discussed the advantages of generic outcome measures over disease- or treatment-specific ones for purposes of outcomes management. These advantages include universality across different types of treatments and the patients' ability to convert generic measures into their personal (and often idiosyncratic) outcomes of interest. These theoretical advantages, important as they may be, would probably be insufficient, in and of themselves, to overcome skepticism concerning the relevance of (what many would see as) a "fuzzy" outcome variable -- global HRQOL.

Standardizing the Concept of HRQOL

One major step toward "de-fuzzifying" HRQOL is to formally limit and standardize, as part of a substantial public education program (see below), what is meant, exactly, by HRQOL in this setting. Respondents must understand that the question(s) being asked concern(s) *health related* quality of life, and that they should specifically exclude from consideration factors beyond the purview of health and health care (e.g., neighborhood crime rate, marital relationships, pollution). What is beyond the purview of health care is, of course, open to discussion; for example, job satisfaction and marital harmony (or lack thereof) might be affected by one's health, or access to health care. Usually, however, such issues should be excluded.

Indeed, for purposes of outcomes management, and as discussed in Part A of this article, it is probably best to restrict health domains in this context to physical suffering and limits on activities.[17] Emotional outlook should be included in HRQOL ratings only to the extent that any emotional problems stem from physical symptoms and activity limits. Otherwise (among other problems), the use of HRQOL as an endpoint would likely prove too "soft" for many physicians and policy-makers. (These remarks assume that physical conditions and treatments are being evaluated. Psychiatric or emotional problems and treatments would, of course, be addressed using different descriptors, a discussion of which is beyond the scope of this article.)

To assist in further standardizing responses to the HRQOL survey, respondents should be asked to rate their average overall HRQOL during the time since the previous survey, e.g., over the preceding six to twelve months -- or, in the case of initial contact with the provider, over the previous month or so. Additional standardization might also be realized by explicitly defining the constructs to be included in the HRQOL ratings (e.g., suffering and activity limits), and by providing definitions for each level of the 11-point scale, perhaps something like this:

Definitions for Quality of Life Survey

Pain and physical symptoms means headache, backache, chest pain, arthritis or any other form of physical pain; also nausea, dizziness, shortness of breath, or any other form of physical suffering. Activity limits include problems with working, housework, hobbies, social relationships, self-care activities (including eating and dressing)

Over the past six months or a year, I have had:

HRQOL
rating **Description**

10. Absolutely no pain or physical symptoms and absolutely no limits on activities

9. Very occasional pain or symptoms of a very minor nature (never need treatment) and/or very minor limits on activities (never restricted in important things)

8. Occasional pain or symptoms, usually of a minor nature (rarely need treatment) and/or minor limits on activities (rarely restricted in important things)

7. Pain or symptoms sometimes troublesome, commonly needs treatment and/or sometimes dissatisfied with extent of limits on activities (sometimes restricted in important things)

And so on. Due to space limitations on the postal card survey, these definitions might best be published in newspapers and information brochures in conjunction with publicizing the outcomes management system, rather than being printed on the survey post-card itself.

Along these lines, a high-profile educational and promotional effort (perhaps sponsored by the government and assisted by the media and health care providers) should be launched to educate the public. both about the nature of HRQOL and, perhaps even more importantly, about the purpose and significance of the outcomes management system. The public should be informed that the survey results will be used to determine "what works" in medicine, and to assist future patients and payers in making decisions about what treatments to undergo or to pay for. Respondents are being asked to contribute their experience to this body of knowledge. This is a weighty responsibility! Indeed, with proper education and promotion about the health outcomes management system and its implications, people could come to view reporting of their HRQOL as a civic duty -- akin to voting in elections. The process of reporting HRQOL could even be incorporated into popular culture ("Hi, how are you?" "Oh, about a 7, I guess.")

In view of this avowed purpose, respondents would be asked to place greater emphasis on symptoms and activity limits that relate to their health care problems and to services for those problems, and to downplay the effects of unrelated or ephemeral health problems. Patients would realize that they were being asked, in effect, "Is the treatment or procedure that you have undergone something you would recommend to others *with your same problem"* and "Should taxpayers pay for that treatment or procedure for patients who can't afford it?" Patients who have not undergone any treatments or procedures for their medical problem would realize that they are being used as a control group to assess health outcomes.

It is hoped that this sort of information will encourage people to participate in the program. Even with the best of efforts, however, many people will fail to respond to the HRQOL questionnaires, in effect refusing to "vote" their HRQOL. This less-than-perfect response will create the potential for a response bias in the data, e.g., the possibility that less-ill (or more-ill patients) will respond disproportionately more (or less) often. It is not clear which way such a bias would work, however, and in any case this sort of democratic effort (one patient-one vote per intervention) must operate based on the votes cast. Those who choose not to vote will be effectively disenfranchised. Of course, every effort should be made to encourage and to facilitate responding. Indeed, this is the very rationale for using the brief postal-card format to solicit HRQOL outcome information.

Induction Problems

Even if global HRQOL is agreed to be a sufficiently "hard" endpoint for large-scale outcome studies, problems remain in ascribing observed changes (or lack of change) in HRQOL to specific treatments and procedures. In Part A we noted that use of global HRQOL measures to derive inferences about treatment outcomes would constitute, in effect, a series of large-scale case-control studies. Actually, though, an even greater causal leap of faith seems to be required than for most case-control studies, which generally focus on specific diseases or conditions, and which evaluate specific factors that possess a fairly tightly posited causal link with those diseases or conditions (for example, the relationship between smoking and lung cancer).

Unlike most case-control studies, in which the outcome is usually a specific condition (e.g., cancer), the outcome of interest in outcomes management is simply the change (if any) in HRQOL over time. Even restricting the definition of HRQOL, as discussed above, such a change (or lack of change) could occur for a large number of reasons, only one of which is the treatment or procedure of interest. For example, patients may sprain their ankle, come down with the flu, or develop an unrelated disease or condition.

Even more troublesome, perhaps, reported changes in quality-of-life might be used to assess the effects of different treatments on the same patient over the same time period. For example, a person who received a cataract operation for decreased vision might report that he changed from a HRQOL of 7 before surgery to a 9 after surgery. When this outcome is used in a study of cataract surgery, the reported improvement in HRQOL would be attributed to the cataract surgery (although the weight assigned to this outcome would be determined by the extent of comorbidities and other factors). If the same patient received a hip replacement operation during the same time period, the very same improvement in HRQOL might be ascribed to the hip surgery during subsequent evaluations of this procedure's effectiveness.

An extensive discussion of the fascinating epistemological problem of inductive inferences[63] is beyond the scope of the present article. Suffice it to say that we must take refuge in large numbers and simply acknowledge the probabilistic nature of the activity which rising costs and concerns about quality have compelled us to undertake. Ascribing observed outcomes (i.e., effects) to antecedent events (i.e., causes) is always a risky business, particularly in an uncontrolled (e.g., retrospective) environment. Nonetheless, the analyses and inferences envisioned for Ellwood-style outcomes management are fundamentally no different than the analyses and inferences regularly relied upon by medical science and policy-makers.

As noted, large numbers of patients are needed in order to strengthen causal inferences emerging from outcomes-management data, and to reliably detect the "signal" of treatment effectiveness amidst the inevitable noise Each arm of outcomes-management case-control studies should probably contain on the order of thousands or tens of thousands of patients. The effect of increasing sample size is directly analogous to increasing the power of a lens in microscopy or astronomy: just as a more powerful lens enables the viewer to resolve a previously indistinct body into its constituent parts, so increasing sample size enables the observer to "magnify" the effects of any given, distinct component of HRQOL.

For example, the change in one hypothetical patient's HRQOL over six months or a year might be divided into the following components:

1. Effect of lung cancer plus chemotherapy
2. Effect of angina plus medical treatment
3. Effect of sprained ankle
4. Effect of arthritis
5. Effect of headache

At the same time, another patient's HRQOL outcome might consist of the following components:

1. Effect of lung cancer without chemotherapy
2. Effect of inflammatory bowel disease
3. Effect of angina plus coronary artery angioplasty

(Note that lung cancer-plus-chemotherapy, and other condition-treatment pairs, should be listed as one item for the first patient; listing, e.g., lung cancer and chemotherapy separately would imply that symptoms deriving from the cancer *per se* are equivalent in both treated and untreated patients -- an unlikely proposition.) If these two patients were included as part of a very large sample of patients with lung cancer, the confounding factors (i.e., all factors except chemotherapy) should be distributed more or less evenly between patients who undergo, versus those who forego, chemotherapy. Again, the "signal" of chemotherapy is magnified and discerned from the "noise" of confounding factors. This effect is related to the well-known statistical tenet that increasing sample size increases the precision of estimates regarding some feature of the population from which the sample is drawn.[64]

From a statistical perspective even the process of ascribing a single HRQOL change (or lack of change) to more than one intervention is perfectly legitimate. On average, the signal representing the HRQOL outcome caused by a particular intervention should be detectable through the noise of all confounding factors -- including other procedures. After all, many more patients receive hip replacements without contemporaneous cataract surgery (and vice versa), and the effects of these relatively few patients is small within the total patient sample. Moreover, both cataract and hip surgery would contribute to any change in HRQOL over the study period, and might produce either additive or counteracting effects on HRQOL. The specific contribution of each procedure can be ascertained by grouping the patient first with a large number of other cataract patients (to be compared with similar patients who did not receive cataract surgery), then (later) with other hip-replacement patients (to be compared with similar patients who did not receive hip replacements).

One problem increased sample size cannot correct is any systematic biases in the data, such as might occur if patients who receive a particular intervention tended to have, say, substantially better social support than those who do not receive that intervention. In this case, if the degree of social support is itself responsible for any favorable outcomes, the analysis will incorrectly ascribe that outcome to the intervention. Analysts and policymakers must be alert for this sort of bias, and should attempt to measure and correct for biasing factors whenever possible. However, the problem of bias in large-scale outcome studies should not be overstated. It seems unlikely that many cases of *meaningful* bias (in the sense that the bias completely invalidates inferences regarding effectiveness) would persist in sample sizes of several thousands or tens of thousands of patients. Moreover, the problem is no different than that faced in the thousands of non-experimental studies already being performed every year. Indeed, the generic problem of bias in observational medical studies would be substantially reduced if researchers had access to the sort of very large data bases envisioned for outcome management programs.

Need for a Central HRQOL Registry

In his original vision of outcomes management,[1] Ellwood suggested that HRQOL information be obtained from patients (only) during encounters with their health care practitioners. Although the initial enrollment of patients is probably best performed during such encounters, subsequent polling should be effected at regular intervals (in order to standardize observed outcomes). It would be next to impossible for practitioners to schedule return visits according to a standard schedule; moreover, patient and practitioner compliance with these data collection efforts would likely be sporadic, at best. For these reasons, follow-up information about HRQOL outcomes should probably be performed via standardized mail surveys, as has been discussed before in these articles.

We propose that a special centralized registry be set up collect health outcome information on all patients within a certain geographic area (e.g., a State). Providers would regularly send to this registry lists of new patients newly diagnosed with one or more of a defined list of serious medical conditions, including cancer, heart disease, arthritis, and the like. Conditions would be selected based on prevalence, severity, cost, and on the existence of substantial variations in how the condition is treated.

Providers would transmit (ideally electronically) the names, mailing addresses, identification numbers, and initial, baseline HRQOL survey results to the central registry, which would then take responsibility for regular mailing of follow-up post-card questionnaires, perhaps every six months or a year. The return questionnaires would be coded (no names) to ensure confidentiality; in addition, strict measures would be adopted by the system to protect patient confidentiality. Registry personnel would take steps to enhance compliance (e.g., reminder postcards or phone calls).

A major function of the registry would be to facilitate linkage of the outcome data with other data bases. This linkage is, of course, what makes outcome data useful -- by permitting inferences to be drawn concerning the effectiveness (or lack of effectiveness) of medical and surgical treatments and procedures. The technology necessary for successfully linking data bases is well-developed. What is lacking, however, is the sort of electronic medical record needed to make full use of the outcome data. Administrative data (e.g., discharge diagnoses, age, gender) can take us a ways down the outcome management road, but, as will be discussed below, there are limitations to this approach.

Perhaps the first successful application of outcomes management will occur in large managed care organizations that have committed to electronic medical records. For example, the Harvard Community Health Plan and the Mayo Clinic provider network are both installing sophisticated electronic medical record systems which, when complete, will cover over 500,000 patients each. A recent article in The New York Times reporting on these data systems recognized their potential role in outcomes management:

> Justifying the costs [of converting to electronic medical records] is a tall order. . . . But the savings could be enormous. The systems could feed information into nation-wide data bases, which could then be studied to determine which treatments work. This invaluable evidence would enable doctors to eliminate unnecessary exams and surgery.[65]

These early efforts, it must be hoped, will serve as harbingers of much bigger things to come. For full-scale implementation and adequate inferential power, even one million patients is probably insufficient, given that each intervention will likely need thousands of patients in each arm. Ideally patients in an entire state or region would be included in the system. Unfortunately, the fragmented payer system characteristic of the United States precludes the sort of large scale population-based patient data base necessary for real outcomes management. Indeed, no population-based data base of the size necessary exists in the United States. For this reason, provinces in Canada would be the ideal laboratories for large scale outcomes management. British Columbia and Manitoba have province-wide data bases on all hospital and physician visits and procedures. Alternatively, a State that succeeds in implementing a single payer system -- rumored to be an option under the coming Clinton health reform package, could attempt such a program. It remains to be seen whether the so-called Health Insurance Purchasing Cooperatives, characteristic of "managed competition" schemes, would be able to assemble the necessary data bases. Clearly, federal leadership (and, perhaps, financing) will be needed if a substantial portion of the United States population is ever to be included within a system of outcomes management.

III. Identifying "Type of Patients"

We now turn to the second major issue pertaining to the interpretation of large-scale observational HRQOL data: identifying patients who either do or do not receive significant benefit from specified treatments and procedures. In a meaningful sense, identifying categories of patients based on differential responses to treatment represents the major task of outcomes analysis and management. Toward this end, it is necessary to classify patients into different "types," based on combinations of clinical and demographic variables. For example, it might be determined that, in order to receive significant benefit from a certain treatment, patients must (1) be over age 65, with (2) mean blood pressure over 120 and (3) serum urea nitrogen under 60. This constellation of indicators denotes a "type of patient."

The predictor variables used to define types of patients are demographic and clinical factors that have been shown to correspond statistically with increased or decreased degrees of benefit, or of some outcome (e.g., mortality). Selection of predictor variables should be guided by studies that have analyzed the relationship of patient outcomes with possible predictors.

Statistical Approaches

There are two basic approaches to this problem: regression techniques and recursive partitioning. Multivariate regression techniques determine the extent of correspondence between each candidate variable and the outcome of interest (e..g, mortality, improved quality of life), holding the effects of other variables constant. These analyses produce equations that provide quantitative estimates of the likelihood of the outcome given any particular set of values on the selected variables. Each variable is assigned a specific numerical weight based on its observed contribution to the outcome; to determine expected outcomes, these weights are multiplied by the values of the variables and the resulting products added to produce the outcome estimate.

Using regression techniques, therefore, "types of patients" would be identified by reference to the probability that they would experience the outcome of interest with treatment. Examples might include "patients who, with treatment, have an expected HRQOL improvement of at least (or less than) 0.5 rating unit compared to no treatment," or "patients who, with treatment, have at least a one-year life-expectancy."

In recursive partitioning,[66] [67] [68] the patient sample is initially divided into two subgroups that are as internally homogeneous as possible with respect to the outcome of interest. This process is repeated recursively within each resulting subgroup until certain stopping rules are invoked. The product of this effort is identification of two or more subgroups, or "types of patients," who differ maximally from each other with respect to the outcome. For example, imagine a sample of cancer patients with an overall mortality rate of 25%, and about whom we know the age, gender, and stage of cancer. Recursive partitioning may determine that, for example, the best "split" on this group is at age 65; perhaps 50% patients older than 65 die, compared to only 10% of patients 65 or younger. If the program chose this split, no other split, whether on age, gender, or stage, could produce greater within-subgroup homogeneity. A second split might then occur in one, both, or neither subgroup; in our hypothetical case, perhaps the older group would be split between less than Stage III cancer versus those Stage III or greater. Either one of these latter subgroups could be split again, either on age or gender. The resulting patient taxonomy ("types of patients") might include 1 (and 2) men (and women) over age 65 with Stage III or greater cancer, (3) patients over age 65 with less than Stage III cancer, and (4) patients aged 65 or younger. The proportion of people in each subgroup with the outcome of interest is known from the HRQOL data base; this proportion becomes the probability that future patients in those subgroups will experience the outcome.

Although regression techniques are generally somewhat more accurate than recursive partitioning,[69] providers would certainly resist the (what they would likely perceive as) an overly quantitative approach to patient classification. The weighted sum (regression) approach is just too reminiscent of "cookbook medicine" i.e., too close to a recipe calling for, say, "3 eggs, 1 cup milk, and 1/2 teaspoon salt." The recursive partitioning approach, on the other hand, produces more-traditional, holistic descriptions of different types of patients based on specified combinations of predictor variables. For this reason, recursive partitioning may be the procedure of choice for identifying patient subgroups who either do or do not derive significant benefit from a particular treatment or procedure.

Much more needs to be said about how, exactly, different types of patients would be identified using the above-described techniques. We are presently developing explicit models of how such process might work, and plan to report our results in the near future.

Clinical versus Statistical Significance

The statistical significance of observed differences in HRQOL between different types of patients can be calculated according to established techniques. Traditional statistical tests must be interpreted with caution, however, because almost any difference across very large numbers of patients will be statistically significant at usual cutoffs (e.g, p < .05 or .01). As noted recently by Diamond and Denton,

> the statistics of hypothesis testing (P values and the associated concept of "statistical significance") were originally designed for sample size around 30, not 30,000. Nevertheless, investigators are now suggesting some questions of clinical interest. . . will require assessment trials involving as many as 140,000 patients. With trials of such size, a clinically inconsequential difference in outcome is readily elevated to the lofty level of statistical significance.[70(p.457)]

With very large sample sizes it might be best simply to dispense with the concept of statistical significance and focus exclusively on clinically meaningful differences in outcomes across differently treated groups. In Part B we suggested that this threshold might be set at one-half rating point on our 0-11 overall HRQOL scale. This difference represents one-twentieth of the distance between best and worst possible HRQOL -- the same fraction commonly accepted as the boundaries of statistical significance (i.e., p <= .05). The choice of a threshold for significant clinical difference is a value judgment and should ultimately be set after public discussion and using due process democratic procedures.[71]

The Problem of Discrimination

Although clinical variables (e.g., blood pressure) are always appropriate variables for defining patient subgroups, demographic variables (e.g., gender, ethnicity) are potentially problematic if the results of effectiveness analyses are to be used for purposes of reimbursement policy or resource allocation. Conclusions to the effect that a treatment "worked" for one demographic group but not another, although perhaps scientifically valid, would be very difficult to accept from a social and political perspective.

> For example, what if a given service is found to provide significant benefit (either on life-expectancy or quality of life) for whites only (or for non-whites only), or for men but not for women (or vice versa)? Would a policy whereby men were covered for, say, chemotherapy for lung cancer, but women were not, be supportable -- even if based on good evidence? . . . The *appearance* of discrimination that would inevitably exist if some but not others were covered for putatively beneficial treatment would be extremely difficult to accept.[5]

Society will need to come to grips with this problem, because it is certain that differences in effectiveness will be observed across demographic lines.

The use of secondary diagnoses and comorbidities could give rise to similar problems. For example, patients who have diabetes may derive significantly less benefit from cardiac transplantation than do patients without diabetes. Although it might be appropriate to counsel diabetic patients about this finding, a formal policy that denied coverage for heart transplants to patients with diabetes might seem discriminatory. If based on good evidence, however, such a policy would probably not be in conflict with the Americans with Disabilities Act or related statutes.[5]

These latter considerations illustrate a problem with the use of administrative data bases that contain only demographic and diagnostic information as potential predictor variables. Lacking clinical data of the type preferred for construction of patient taxonomies, administrative data bases can only provide limited degrees of useful insight into the question of "who benefits" from a particular treatment. However, administrative data can provide information concerning differences in outcomes based on which provider is used. This type of data is becoming of increasing significance in this era of quality assessment[72] and "outcome report cards."[73]

There is another important methodological consideration regarding the identification of patient subgroups. It is well known that retrospective searches for subgroups inevitably "succeed" in finding some types of patients who appear to respond to a given intervention even when the entire sample, taken as a whole, does not respond.[74] Often such findings are due strictly to chance, the result of "data dredging." Indeed, it is statistically certain that obviously nonsensical subgroups will be found to benefit from treatment, perhaps patients who are Scorpios and have last names beginning with a "T."

In general, patient subgroups who are determined to benefit from treatment should pass two tests before the results are accepted and acted upon. First, the effect should persist across different States or other large subsets of the population, i.e., it must be reproducible. Second, the finding must make clinical sense; there must be some more-or less accepted clinical theory that can account for the finding. In borderline cases, the findings should be tested in prospective studies, perhaps using longer generic HRQOL questionnaires (such as the SF-36) and disease- or treatment-specific measures. Some procedural mechanism, perhaps a standing Commission, will be needed to determine which findings should be taken at face value, which rejected, and which subjected to prospective study.

Length of Life versus Quality of Life

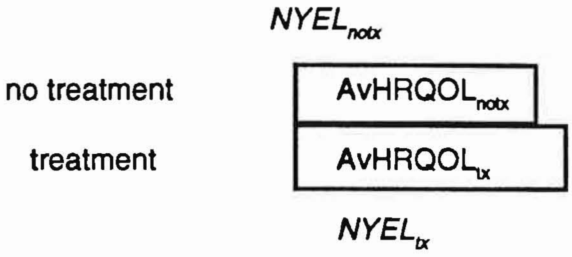
As a final consideration in developing taxonomies of "types of patients," it must be noted that in order to properly evaluate treatment effectiveness it is necessary to consider not only HRQOL outcomes, but also the impacts of treatment (if any) on life-expectancy. Some treatments, such as an appendectomy for appendicitis, are valued not for any improvements in quality of life, but rather for their substantial effects on life-expectancy (e.g., in the case of appendicitis, from perhaps two weeks untreated to normal life-expectancy given appropriate treatment). Other treatments have significant impacts on quality of life, but little or no effect on life-expectancy -- such as medication or surgery for arthritis, or prostatectomy for benign obstruction. Still other treatments involve trade-offs between quality and quantity of life, where a longer life-expectancy comes at the expense of various symptoms resulting in a possible decrease in quality of life. Chemotherapy for some forms of cancer might fall into this category.

The basic model for evaluating treatment effectiveness can be summed up in algebraic terms:

$$\text{Effectiveness} = [(\text{AvQOL}_{tx} \times \text{NYEL}_{tx}) - (\text{AvQOL}_{notx}) \times (\text{NYEL}_{notx})]$$

AvHRQOL_{tx} = Average HRQOL with treatment (scaled between 0-1)
NYEL_{tx} = Number of Years Expected to Life with treatment
AvHRQOL_{notx}) = Average HRQOL without treatment
NYEL_{notx} = Number of Years Expected to Life without treatment

These terms can be depicted in graphical form:



The difference in "weight" (quality x quantity) between the top and bottom boxes constitutes the net effectiveness of the treatment. Note that the weight of the bottom (treatment) box may be less than that of the top box if treatment results in overall lower life expectancy and/or worse HRQOL.

In the end, the goal of outcomes management is to identify types of patients whose "treatment boxes" "weigh" significantly more than their "no-treatment boxes" with respect to a specific intervention. The extent to which the treatment box should outweigh the non-treatment box before the difference is considered "significant" -- and the intervention deemed effective -- constitutes a value judgment which, again, calls for public debate and the use of democratic evidence-based decision-making procedures.

IV. Conclusion

With due care, the collection and analysis of global HRQOL outcome data can provide valuable information concerning the effectiveness of medical treatments and procedures. Reporting of one's health outcomes should come to be viewed as a civic responsibility, encouraged by providers and policymakers. Large data sets must be assembled. Appropriate data collection infrastructures should be developed as soon as possible, and analysis and interpretation issues debated and resolved.

The festering problems of increasing health care costs and inequitable access to effective services demand that large scale outcomes measurement and management systems, as suggested by Ellwood over five years ago, become a reality.

REFERENCES

1. Ellwood PM. Outcomes management: a technology of patient experience. N Engl J Med 1988; 318: 1549-56.

2. Epstein AM. The outcomes movement: will it take us where we want to go? N Engl J Med 1991.

3. Relman AS. Assessment and accountability: the third revolution in medical care. N Engl J Med 1988; 319: 1220-2.

4. Wennberg JE. Outcomes research, cost containment, and the fear of health care rationing. N Engl J Med 1990; 323: 1202-4.

5. Hadorn DC. The problem of discrimination in health care priority setting. JAMA 1992; 268: 1454-1459.

6. Hadorn DC. Setting health care priorities in Oregon: cost-effectiveness meets the Rule of Rescue. JAMA 1991; 265: 2218-2225.

7. Eddy DM. Oregon's plan: should it be approved? JAMA 1991; 266: 2439-45.

8. Fox DM, Leichter HM. Rationing care in Oregon: the new accountability. Health Affairs 1991 10:2:7-27.

9. Gareiss R. Beyond billing. Am Med News, January 25, 1993, pp. 9-10.

10. Deyo RA, Carter WB. Strategies for improving and expanding the application of health status measures in clinical settings. Med Care 1992 Supplement; 30: MS176-MS186.

11. Lansky D, Butler JBV, Waller FT. Using health status measures in the hospital setting: from acute care to 'outcomes management.' Med Care 1992; 30 *Supplement*: MS57-MS73.

12. Campbell DT, Stanley JC. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1966.

13. Sechrest L, Perrin E, Bunker J., (eds.) Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. Rockville, MD: United States Department of Health and Human Services. Agency for Health Care Policy and Research, 1991.

14. Kaplan RM and Bush JW. Health-related quality of life measurement for evaluation research and policy analysis. Health Psychology 1982; 1: 61-80.

15. Patrick DL and Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Medical Care 1989; 27 *Supplement*: S217-S232.

16. Ware JE. Methodological Considerations in the Selection of Health Status Assessment Procedures. In Wenger, NK, Mattson ME, Furberg CD, Elinson J (eds.), Assessment of Quality of Life in the Clinical Trials of Cardiovascular Therapies. New York: LeJacq Publishing, Inc., 1984, pp. 87-111.

17. Hadorn DC. The role of public values in setting health care priorities. Soc Sci Med 1991; 32: 773-782.

18. Bush JW, Fanshel S, Chen M. Analysis of a tuberculin testing program using a health status index. Social-Economic Planning Sciences 1972; 6: 49-69.

19. Bush JW, Chen M, 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: Hospital Research and Educational Trust, 1973, p. 172-208.

20. Stewart A.L., Greenfield S., Hays R.D., et al. Functional status and well-being of patients with chronic conditions. JAMA 262, 907-913, 1989.

21. McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires. New York: Oxford University Press, 1987.

22. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36):I. conceptual framework and item selection. Med Care 1992 Vol. 30 pp. 473-483.

23. Inter-Patient Outcome Research Team's Outcomes Assessment Work Group Survey of Outcomes Assessment Instruments Used in Primary Data Collection and Data Collection Instruments. Alexandria, VA: Walcoff, 1992.

24. Kaplan RM, Anderson JP. The general health policy model: an integrated approach. In Quality of Life Assessments in Clinical Trials (Spilker B, ed.). New York: Raven Press, 1990.

25. Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. J Roy Coll Gen Pract 1985; 35: 185-89.

26. McKenna SP, Hunt SM, McEwen J. Weighting the seriousness of perceived health problems using Thurstone's method of paired comparisons. Int J Epidem 1981; 10: 93-97.

27. Bergner M, Bobbitt RA, Carter WB, et al. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981; 19: 787-805.

28. Rosser R, Kind P. A scale of valuations of states of illness: is there a social consensus? Int J Epidem 1978; 7: 347-58.

29. Hadorn DC, Hays RD. Multitrait-multimethod analysis of health-related quality of life measures. Med Care 1991; 29: 829-840.

30. Hadorn DC, Hays RD, Uebersax J, et al. Improving task comprehension in the measurement of health state preferences: a trial of informational figures and a paired comparison task. J Clin Epi 1992; 45: 233-243.

31. Analysis Under the Americans with Disabilities Act (ADA) of the Oregon Reform Demonstration. Washington, DC: US Department of Health and Human Services; August 3, 1992.

32. Najman J, Levine S. Evaluating the impact of medical care and technologies on the quality of life: A review and critique. Soc Sci Med 1981; 107-115.

33. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. J Chron Dis 1978; 31: 697-704.

34. Slevin ML, Stubbs L, Plant HJ, et al. Attitude to chemotherapy: comparing views of patients with cancer with those of doctors, nurses, and general public. BMJ 1990; 300: 1458-1460.

35. Christensen-Szalanski JJJ. Discount functions and the measurement of patients' values: Women's decisions during childbirth. Med Dec Making 1984; 4: 47-58.

36. O'Connor AM, Boyd N, Warde P. Eliciting preferences for alternative drug therapies in oncology: Influence of treatment outcome description, elicitation technique and treatment experience on preferences. J Chron Dis 1987; 40: 811-818.

37. LLewellyn-Thomas HA, Sutherland HJ, Thiel EC. Do patients' evaluations of a future health state change when they actually enter that state? Med Decis Making 1991:11:323 (abstract).

38. Kaplan RM, Bush JW, Berry CC. The reliability, stability, and generalizability of a health status index. American Statistical Association, Proceedings of the Social Status Section, 1978, 704-709.

39. Froberg DG, Kane RL. Methodology for measuring health-state preferences--III: population and context effects. J Clin Epidemiol 1989; 42: 585-92.

40. Rokeach M. The Nature of Human Values. New York: The Free Press, 1973.

41. Balaban DJ, Sagi PC, Goldfarb NI, et al. Weights for scoring the quality of well-being instrument among rheumatoid arthritics: A comparison to the general population weights. Med Care 1986; 24: 973-980.

42. LLewellyn-Thomas H, Sutherland HJ, Tibshirani R, et al. Describing health states: Methodological issues in obtaining values for health states. Med Care 1984; 22: 543-552.

43. Boyle MH, Torrance GW. Developing multiattribute health indexes. Med Care 1984; 22: 1045-1057.

44. Thurstone L. A law of comparative judgment. Psychological Review 1927; 34: 273-286.

45. Fylkesnes K, Forde OH. Determinants and dimensions involved in self-evaluation of health. Soc Sci Med 1992; 35: 271-279.

46. Barsky AJ, Cleary PD, Klerman GL. Determinants of perceived health status of medical outpatients. Soc Sci Med 1992; 34: 1147-1154.

47. Campbell DT, Fiske DW. Converent and discriminant validation by the multitrait-multimethod matrix. Psych Bull 1959; 56: 81-105.

48. Guyatt, G., Deyo, R.A., Charlson, M., et al. Responsivness and validity in health status measurement: A clarification. J Clin Epidemiol 1989:42:403-408.

49. Hays RD, Hadorn DC. Responsiveness to change: an aspect of validity, not a separate dimension. Quality of Life Res 1992; 1: 73-75.

50. Reuben DB, Rubenstein LV, Hirsch SH, Hays RD. Value of functional status as a predictor of mortality: results of a prospective study. Am J Med 1992; 93: 663-9.

51. Wolinsky FD, Johnson RJ. Perceived health status and mortality among older men and women. J Gerontol 1992; 47: S304-312.

52. Reuben DB; Siu AL; Kimpau S. The predictive validity of self-report and performance-based measures of function and health. J Gerontol 1992; 47: M106-110.

53. Shepherd SL, Hovell MF; Slymen DJ; Harwood IR; Hofstetter CR; Granger LE; Kaplan RM. Functional status as an overall measure of health in adults with cystic fibrosis: further validation of a generic health measure. J Clin Epidem 1992; 45: 117-125.

54. Mossey JM, Shapiro E. Self-rated health: A predictor of mortality among the elderly. AJPH 1982; 72: 800-808.

55. Gough IR, Furnival CM, Schilder L, et. al. Assessment of the quality of life of patients with advanced cancer. Eur J Cancer Clin Oncol 1983; 19: 1161-5.

56. Feinstein AR. Benefits and obstacles for development of health status assessment measures in clinical settings. Med Care 1992; 30 *Supplement*: MS50-MS56.

57. Andrews FM, Crandall R. The validity of measures of self-reported well-being. Soc Indicat Res 1976; 3: 1-19.

58. Andrews FM, Withey SB. Social indicators of well-being: Americans' perceptions of life quality. New York: Plenum, 1976.

59. Lehman AF, Ward NC, Linn Ls. Chronic mental patients: the quality of life issue. Am J Psychiatry 1982; 139: 1271-1276.

60. Atkinson T. The stability and validity of quality of life measures. Soc Indicat Res 1982; 10: 113-132.

61. Campbell DT, Stanley JC. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1966.

62. Sechrest L, Perrin E, Bunker J., (eds.) Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. Rockville, MD: United States Department of Health and Human Services. Agency for Health Care Policy and Research, 1991.

63. Weed DL. On the logic of causal inference. Am J Epidem 1986; 123: 965-979.

64. Wonnacott RJ, Wonnacott TH. Introductory Statistics (4th edition). New York: John Wiley, 1985.

65. Rifkin G. New momentum for electronic patient records. New York Times, May 2, 1993, F8.

66. Marshall RJ. Partitioning methods for classification and decision making in medicine. Stat Med 1986; 5: 517-526.

67. Cook EF, Goldman L. Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis. J Chron Dis 1984; 37: 721-731.

68. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Belmont, CA: Wadsworth International Group, 1984.

69. Hadorn DC, Draper D, Rogers WH, Keeler EB, Brook RH. Cross-validation performance of mortality prediction models. Stat Med 1991; 11: 475-489.

70. Diamond GA, Denton TA. Alternative perspectives on the biased foundations of medical technology assessment. Ann Int Med 1993; 118: 455-464.

71. Hadorn DC. Emerging parallels in the American health care and legal- judicial systems. Am J Law Med 1992; 18: 73-96.

72. Darby M. Minnesota blues re-allocates resources using outcomes. Med Guidelines and Outcomes Res 1992; 3: 1-2.

73. Mitka M. Ready or not, here come outcomes 'report cards'. Am Med News March 22/29, 1993: 4.

74. Bulpitt CJ. Subgroup analysis. Lancet July 2, 1988; 31-34.