Title
A critique of current uses of health status for the assessment of
treatment effectiveness and quality of care.

Permalink
https://escholarship.org/uc/item/6t71h7rs

Journal
Medical care, 38(9 Suppl)

ISSN
0025-7079

Authors
Kaplan, SH
Kravitz, RL
Greenfield, S

Publication Date
2000-09-01

License
CC BY 4.0

Peer reviewed
A Critique of Current Uses of Health Status for the Assessment of Treatment Effectiveness and Quality of Care

SHERRIE H. KAPLAN, MPH, PHD,* RICHARD L. KRAVITZ, MD, MSPH,† AND SHELDON GREENFIELD, MD*

The great advances in the measurement of health-related quality of life (HRQOL) have yielded numerous brief, precise, and psychometrically sound instruments. However, despite their widespread use in the relatively narrow arena of randomized controlled clinical trials,¹-⁴ a strong and consistent relationship between patient-reported health status and health care has yet to be demonstrated. In fact, the applications of general health status measures to the evaluation of the “effectiveness” and “quality” of medical care have yielded a mixed and confusing picture.⁵-⁸ There have been few directional hypothesis-driven empirical studies that declare and evaluate the mechanism through which the process of medical care might affect change in HRQOL.³ If not addressed, this absence of theoretically directed research will continue to restrict its use to research, without diffusion into practice. The core thesis of our comments on the excellent and carefully thought out report by Testa et al⁹ in this issue of Medical Care is that unless there are clinically plausible and a priori specified links between medical care and HRQOL in a given study, any observed associations are likely to be spurious. It is one thing to assess the health of a community or population, the purpose for which general health status measures were originally intended. It is very different problem to investigate the relationship of an individual’s perception of his or her health to structural features or specific processes of medical care.

Interpretation of HRQOL Study Results

With respect to “minimally important differences” in HRQOL, the perspective of the user of effective-ness or quality of care data directly affects the magnitude of these differences. For the patient, for example, the minimally important difference is the increment in health status that is “noticeable” as improvement or worsening. For the clinician, important differences are the amount of change in health status measures that would warrant a change in a patient’s or group’s treatment plan. For the policy maker, minimally important health status differences may be those unconfounded differences between populations who are cared for by different types of providers or different systems of care that would warrant a policy change, such as limiting HIV care to HIV specialists or subsidizing the primary care training of family physicians over general internists. It could be argued that for policy purposes, the magnitude of observed health status differences needs to be quite large due to the array of competing explanations for and extraneous variables that contribute to differences in HRQOL between populations.

The only way to assess the magnitude of meaningful health status differences for use in applied health policy purposes is to focus on specific, clinically plausible hypotheses. Suppose the hypothetical research findings in Table 1 were observed between 2 groups of patients undergoing a surgical procedure in different systems of care. In each case, there is a 10-point net difference between health maintenance organizations (HMOs) and fee-for-service (FFS) organizations. Even with the assumption that the groups are otherwise equal with respect to patient characteristics, which, if any, differences are meaningful? We could argue that true variations in care would themselves more directly affect pain than physical functioning and

*From the New England Medical Center, Primary Care Outcomes Research Institute, Boston, Massachusetts.
†From the University of California, Davis, Center for Health Services Research in Primary Care, Sacramento, California.

Address correspondence to: Sherrie H. Kaplan, MPH, PhD, Primary Care Outcomes Research Institute, 750 Washington St, NEMC #345, Boston, MA 02111. E-mail: skaplan1@lifespan.org
more directly affect physical functioning than social functioning because of the plausible conceptual distance of these outcomes from the surgical procedure in the causal chain. From a policy perspective, we should accept the findings on pain and physical functioning (favoring FFS) more readily than the findings on social functioning (favoring HMOs). This example illustrates the problem in affixing a uniform minimal interpretable difference standard across all dimensions of health status even for the same medical care situation, even assuming the groups are similar, which they often are not. The next problem comes in quantifying the appropriate magnitude of HRQOL differences from a policy perspective. We address this problem later.

Two assumptions underlie the development of hypotheses and the definitions of meaningful differences. The first and most important of these assumptions is that medical care antedates HRQOL “outcomes.” Building evidence suggests that the observed stability in general health status may be rooted in its connection to more basic and relatively immutable characteristics of the individual, such as socioeconomic status, genetic background, or personal history. If patients’ perceptions of their general health status are based on a lifetime of experience with their ability to function in their physical, emotional, and social environment, helped or hindered by the medical care they have received, the societal support they have garnered, the environment in which they have been surrounded, and their self-defined standards of that performance, then relatively minor adjustments in a specific medical regimen could not be expected to move these perceptions in any meaningful way. On the other hand, the diagnosis of a new and devastating disease, a major surgical procedure, or the introduction of a complex regimen with debilitating side effects could be expected to influence these perceptions, at least temporarily. The unchallenged assumption that medical care antedates HRQOL must be examined in each research circumstance.

Second, characteristics of both the individual and the environment simultaneously influence biological and physiologic variables, as well as symptom status, functional status, general health perceptions, and overall quality of life. The relationship between quality of medical care HRQOL outcomes may be confounded by extraneous variables in unexplained ways all along the posited causal chain. Furthermore, the problem becomes much worse as the distance increases between the implemented process of medical care and the outcomes that the care is hypothesized to affect.

For example, with the assumption that adherence to published guidelines on the use of combination hypoglycemic therapy in type 2 diabetes mellitus improves glycemic control, most models would generate the causal chain shown in Fig. 1. With the assumption that the quality of medical care (use of oral agents) accounts for 50% of the variance in glycemic control (a very powerful intervention, indeed), all other characteristics of individuals and the environment would account for the substantial remaining variation. With the further assumption that enhanced glycemic control accounts for 50% of the variance in self-reported polyuria, which in turn accounts for 50% of the variance in sleep and energy, and so on, the percentage of variance in overall quality of life accounted for by better use of hypoglycemic agents would be \((0.5)^6 = 1.56\%\). This is the estimate of variance that would accrue to treatment as a lower limit because it is possible that better use of hypoglycemic agents would indirectly affect quality of life, via other causal mechanisms.

Nevertheless, the example highlights how the strength of the relationship between effectiveness or quality of care and outcomes declines exponentially the further removed the outcome is from a direct relationship to medical care. Even when 2 variables (eg, diabetes clinical severity and physical functioning) are spaced fairly closely together on the causal chain, substantial unexplained variance remains. In patients with diabetes, the severity of diabetes and the severity of associated comorbidities, as measured with the Total Illness Burden Index, accounted for \(\approx 25\%\) of the variance in the 10-item physical functioning dimension of the SF-36. The problem when the outcome of interest and medical care are increasingly separated, as shown in Fig. 2, becomes much worse. Brenner et

### Table 1. Hypothetical Changes in HRQOL Among Patients Undergoing Laminectomy in FFS vs HMO

<table>
<thead>
<tr>
<th>Plan Type</th>
<th>Pain</th>
<th>Physical Functioning</th>
<th>Social Functioning</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMO</td>
<td>+10</td>
<td>+10</td>
<td>+20</td>
</tr>
<tr>
<td>FFS</td>
<td>+20</td>
<td>+20</td>
<td>+10</td>
</tr>
</tbody>
</table>

Vol. 38, No. 9, Suppl. II HEALTH STATUS, EFFECTIVENESS, AND QUALITY
al used the case of cataract surgery to highlight this issue.

We propose that a patient's HRQOL, as measured with current instruments, is likely to be modified by medical care in only very discrete, definable subdimensions and then only modestly. It is unrealistic to assume, for example, that measures like the SF-36 would or should be useful in detecting the effectiveness of such drugs as Viagra. Why should those who respond favorably in clinical terms to the intended use of this drug be able to walk more blocks? The current measures of HRQOL are clearly those that matter at least as much as clinical endpoints to patients. On the other hand, the measurement community has been slow to accept that in most instances, medical care has only an indirect and a rather small effect on such outcomes. Therefore, to conduct fair and accurate assessments of the effectiveness and quality of medical care, we should be principally measuring the endpoints that medical care is specifically and directly designed to affect. Unfortunately, such measures are primarily disease or condition specific. We also need to measure “ultimate” outcomes of care like HRQOL. If the 2 sets of outcomes move in different directions, however, we should be very cautious about accepting the HRQOL results. That is, the threshold for meaningful differences for policy and clinical relevance should be even higher.

**Calibration of HRQOL**

These considerations affect the calibration and therefore the interpretation of the meaningfulness of observed differences in HRQOL. Testa argues that some clinical anchor is key to the interpretation of HRQOL differences. We concur with this assertion. However, we regard as inadequate the efforts by others to calibrate HRQOL measures using statistical tests or such variables as mortality, the Holmes-Rahe Stressful Life Events Scale, or patient-reported global health perception measures. The use of SEM or mean differences or effect sizes does not contribute to calibration, because they involve only the variability in the HRQOL measure itself. The use of variables such as the Holmes-Rahe scale is problematic for several reasons. Apart from the possible calibration of HRQOL for use in the evaluation of mental health care, the use of such a measure has narrow clinical care application. Furthermore, the Holmes-Rahe measure itself has long been noted to have serious measurement error problems, nonreplicable weights, and instability over short time intervals. The original weightings were obtained from veterans of World War II, a highly stressful event. Compared with such an event, apart from the death of a spouse, other events represented a relatively random array at the bottom of the continuum. Weights did not replicate well in more
diverse populations. Interpretation of relationships between such measures and HRQOL, unsupported by clinical plausibility, makes high the likelihood of the acceptance of spurious associations as meaningful.

The use of mortality as a calibrating variable is also problematic for several reasons. It is often substantially distant in time from the more proximate variations in HRQOL. It is rare in more proximal clinical circumstances and settings. It also lacks specificity in relation to the majority of clinical care. The use of a patient-reported global measure of health perceptions as a calibrating variable, as advocated by some researchers, presents a tautology, especially in the evaluation of effectiveness and quality of medical care. Global patient ratings of specific patient-reported health status items (as in HRQOL measures) yields only internal construct validity (ie, confidence that internal to the HRQOL concept, we are measuring what we think we are measuring).

What is left? In most instances, what is left for the reasonable calibration of HRQOL measures for use in the evaluation of effectiveness and quality of medical care are the variables that would be considered most important and most credible in clinical settings (ie, clinical measures). These clinical measures, which represent clinical states, have 2 virtues: (1) they are credible to and can be interpreted by physicians and other health care providers, patients, representatives of health care delivery systems, third party payors, and policy makers; and (2) they are causally related to at least some of the HRQOL in the clinical setting. Although Testa supports the use of clinical measures to calibrate HRQOL measures, she goes on to assert that the measurement properties of patient-reported HRQOL measures are more “subjective”
and less sound than clinical measures. We do not concur with this assertion.

The measurement properties of clinical measures are not more sound than properties of patient-reported measures. Thyroid function, liver function, renal function, cardiac function, and even blood sugar are abstract complex concepts. Individual measures of these concepts are subject to multiple sources of measurement error for which they are often not tested. Many clinical measures, such as radiographs, biopsy slides, and so on, that rely on expert interpretation have substantial problems with interrater reliability, some of which have only recently been empirically assessed. Other measures that involve assays, for example, tend to have poor measurement properties over the entire range of values and can have interlaboratory instability. Even the presence or absence of disease (eg, depression versus no depression, myocardial infarction versus no myocardial infarction, as mentioned by Testa) may be inaccurately reported, itself dependent on the aggregation of clinical measures, each with associated errors. Furthermore, such measures are at best crude indicators of the extent of illness and therefore may provide limited help with the calibration of HRQOL.

In practice, clinicians rarely use values of individual clinical measures alone in patient care. Testa gives an example of a clearly important difference in an individual clinical measure: a change from a fasting blood glucose value of 180 mg/dL to 100 mg/dL. But what can be said about the clinical importance of a 10% difference in this value, say from 180 mg/dL to 160 mg/dL? How important is such a difference in terms of future retinopathy, nephropathy, or neuropathy, all surrogate markers of clinically severe diabetes? In fact, blood glucose differences or any differences in the majority of individual clinical measures at a point in time may not be easier to interpret than HRQOL measures. The use of individual clinical measures to calibrate HRQOL measures and to establish acceptable “meaningful” category intervals therefore becomes highly problematic.

Perhaps a more appropriate approach to the evaluation of the shared variation between clinical and HRQOL constructs would be to compare multi-item aggregations of both. Although not currently part of the assessment of liver function, alkaline phosphates, bilirubin, SGOT, and SGPT might be best combined into a scale, according to the aggregation principles used in cluster analyses, such as factor analysis. Weighting could be clinically defined, driven by the outcome of interest (eg, diagnosis of primary biliary cirrhosis versus acute hepatitis). Even blood sugar, a less complicated concept than liver function, is represented by

<table>
<thead>
<tr>
<th>Disease/Condition</th>
<th>Least</th>
<th>Minimal</th>
<th>Moderate</th>
<th>Severe</th>
<th>F*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hearing loss</td>
<td>62.9</td>
<td>51.4</td>
<td>38.9</td>
<td>43.3</td>
<td>19.8</td>
</tr>
<tr>
<td>Hypertension</td>
<td>59.1</td>
<td>61.9</td>
<td>55.6</td>
<td>46.5</td>
<td>6.1</td>
</tr>
<tr>
<td>Nonspecific bowel disease</td>
<td>62.7</td>
<td>56.4</td>
<td>41.4</td>
<td>33.3</td>
<td>25.2</td>
</tr>
<tr>
<td>Genitourinary problems</td>
<td>62.9</td>
<td>51.9</td>
<td>40.8</td>
<td>39.4</td>
<td>24.2</td>
</tr>
<tr>
<td>Gastrointestinal autonomic neuropathy</td>
<td>61.3</td>
<td>36.8</td>
<td>36.2</td>
<td>31.1</td>
<td>31.7</td>
</tr>
<tr>
<td>Foot disease</td>
<td>69.6</td>
<td>55.9</td>
<td>36.6</td>
<td>27.0</td>
<td>63.7</td>
</tr>
<tr>
<td>Lower gastrointestinal disease</td>
<td>62.9</td>
<td>46.6</td>
<td>40.6</td>
<td>37.5</td>
<td>24.8</td>
</tr>
<tr>
<td>Upper gastrointestinal disease</td>
<td>60.8</td>
<td>. . .</td>
<td>49.3</td>
<td>38.0</td>
<td>17.3</td>
</tr>
<tr>
<td>Musculoskeletal problems</td>
<td>64.5</td>
<td>. . .</td>
<td>46.7</td>
<td>37.6</td>
<td>44.6</td>
</tr>
<tr>
<td>Vision problems</td>
<td>65.8</td>
<td>50.9</td>
<td>43.4</td>
<td>39.0</td>
<td>31.9</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>67.8</td>
<td>. . .</td>
<td>46.3</td>
<td>36.1</td>
<td>63.7</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td>66.2</td>
<td>49.5</td>
<td>35.8</td>
<td>27.0</td>
<td>51.7</td>
</tr>
<tr>
<td>Ischemic heart disease</td>
<td>62.5</td>
<td>50.8</td>
<td>40.0</td>
<td>32.3</td>
<td>27.6</td>
</tr>
<tr>
<td>Renal disease</td>
<td>58.8</td>
<td>. . .</td>
<td>42.6</td>
<td>29.5</td>
<td>18.5</td>
</tr>
<tr>
<td>Neurological problems</td>
<td>60.8</td>
<td>41.7</td>
<td>33.7</td>
<td>21.1</td>
<td>27.7</td>
</tr>
</tbody>
</table>

Values are mean, adjusted for age, gender, and income (*P <0.001).
Reprinted with permission from Greenfield et al.18
a number of clinical measures, such as hemoglobin A1c, random blood sugar, 2-hour postprandial blood sugar, fasting blood sugar, 24-hour urine sugar, and so on. Although each of these shares substantial common variance, they also clearly measure different aspects of the concept that could be labeled “blood sugar.”31 A composite could theoretically be a better measure of circulating blood sugar than any one measure used individually.

**Adjustment of HRQOL**

For this reason, we and others have used multiattribute-scaling techniques to construct aggregate measures of clinical states. For example, a moderately elevated random or fasting blood glucose (or hemoglobin A1c), in conjunction with fatigue and polyuria, would be used in combination to indicate a hyperglycemic state sufficiently high to warrant treatment initiation or change; a single measure of blood sugar by itself might not induce such a change. To adjust HRQOL for baseline case mix, we developed a measure of the patient’s total illness burden (the Total Illness Burden Index [TIBI]) that combines aggregate measures of patient-reported disease severity for each of 15 diseases.18–21 For lung disease, for example, there are 5 scaled items: severity of shortness of breath, cough, episodes of bronchitis, presence of 3 common lung conditions, and number of pillows slept on at night. These were aggregated through the application of clinical definitions of disease severity into a scale that ranged from 5 to 20. This scale was then divided empirically into 4 severity levels (none, mild, moderate, and severe) for clinical sensibility. All severity scales were individually tested against clinically relevant dimensions of the SF-36.

All 15 dimensions were then weighted by cluster analysis. The aggregate scale combined ~100 items aggregated across 15 dimensions. Each severity dimension can be weighted differently depending on the index condition under study. For example, cardiovascular disease, eye disease, and kidney disease may receive higher weights in the aggregate TIBI when diabetes is the index condition under study than when the index condition is prostate disease. We used the TIBI to calibrate relevant subdimensions of the SF-36 in a variety of studies of the effectiveness of medical care. For example, in Table 2, we show that each level of lung disease is different from the next by >50% of an SD in the role function due to physical health scale of the SF-36 (fourth line from the bottom). Differences of similar magnitudes were found for the remainder of the organ system variables. These differences represent meaningful clinical differences to both doctor and patient that can be defined in very specific terms, such as having to sleep on an extra pillow at night, producing 2 additional tablespoons of sputum, or being unable to walk a second flight of stairs without becoming short of breath.

Table 3 shows that when these individual dimensions of disease severity are themselves aggregated into the TIBI, they preserve a linear relationship to HRQOL measures. Each severity level of the TIBI was different from the next by at least 30% to 50% of an SD. To verify that these relationships were not limited exclusively to HRQOL but included other indicators of health status, we also related TIBI to the use of health care services, to days lost from work or usual activities (Table 3), and to mortality rates in other studies. Findings from these studies confirmed the validity of the TIBI.

---

**Table 3. The Relationship of Comorbidities to SF-36 Domains in Patients With Prostatic Cancer**

<table>
<thead>
<tr>
<th>SF-36 Domain</th>
<th>Burden of Comorbidities (TIBI)</th>
<th>P, F Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical function*</td>
<td>Least 92.2</td>
<td>Moderate 79.1</td>
</tr>
<tr>
<td>Bodily pain*</td>
<td>Least 91.7</td>
<td>Moderate 73.2</td>
</tr>
<tr>
<td>Mental health*</td>
<td>Least 85.9</td>
<td>Moderate 77.9</td>
</tr>
<tr>
<td>General health perception*</td>
<td>Least 81.0</td>
<td>Moderate 61.6</td>
</tr>
<tr>
<td>No. of bed days (3 mo)</td>
<td>Least 0.00</td>
<td>Moderate 0.16</td>
</tr>
</tbody>
</table>

*SF-36 domains scored from 0 to 100, with 100 = optimal HRQOL.

Reprinted from *Urology*, Vol. 54, Steir et al.,20 with permission from Elsevier Science.
It should be noted that because the very strength of measures such as the TIBI is that a given score can be achieved in numerous ways, we have not advocated the use of composite clinical measures as outcomes of care. We have restricted their uses to baseline case mix assessment, validation of health status measures, or assessment of the clinical health of a population.

Guidelines for the Use of HRQOL in Clinical Studies

The same guidelines that apply to causal inference in epidemiology or in any research discipline should also apply in outcomes research. To infer that medical care is responsible for an improvement in HRQOL, the following should be present: (1) observed differences should meet or exceed modest effect sizes for the outcome variables (eg, 0.40 to 0.60); (2) a dose-response relationship should exist between care and outcomes (ie, more care, better care, more intensive services, or a more pure organizational delivery model leads to better and better outcomes); (3) temporality must be met (ie, a change in delivery of care precedes a change in outcomes in time); and (4) observed differences must be biologically or clinically plausible (as specified in a priori hypotheses). In a review of numerous studies in which HRQOL was used as an outcome, we found that these guidelines were relatively commonly violated:

1. Explicit a priori directional hypotheses should be specified along with power, specification of meaningful differences in HRQOL, and so on.
2. Studies should be designed with HRQOL as the primary endpoint.
3. Clinical endpoints should be measured coincidently with medical care or in plausible temporal relation to the delivery of medical care.
4. Results should be interpreted based on plausibility; the strength and magnitude of the hypothesized association should be sensible, and consistency with past literature or dose response should be characterized.
5. Post hoc subgroup analyses (type 1 error) should be avoided.
6. The sensitivity or specificity of the HRQOL measures to better medical care should be supported.
7. Time intervals for the assessment of QOL (temporality) should be specified or, if not specified, theoretically justified.
8. HRQOL subscales should not be aggregated without justification for the aggregation (ie, were all dimensions expected to be equally affected by care and in the same direction); the impact of aggregation on the relationship to medical care should be specified.
9. An adjustment for baseline differences in case mix (selection bias and confounding) should be performed routinely.

Inattention to any of these basic threats to validity makes any observed relationship between health or medical care and HRQOL very likely to be uninterpretable.

In summary, the calibration of specific dimensions of the HRQOL can be accomplished in a meaningful way, leading to interpretable differences, with the use of aggregated clinical measures, combined with careful case-mix adjustments, and the creation of plausible hypotheses that link clinical measures with HRQOL. Until calibration studies that link HRQOL and clinical measures are replicated in a broad array of clinical circumstances and settings, these measures are best used in combination to evaluate the effectiveness and quality of medical care.

Acknowledgment

The authors gratefully acknowledge the critical review and comment in the development of the manuscript by Grace A. Connolly, JD.

References

5. Mark DB, Naylor CD, Hlatky MA, Califf RM, Topol EJ, Granger CB, et al. Use of medical resources and

6. Ware JE, Bayliss MS, Rogers WH, Kosinski M, Tarlov AR. Differences in 4-year health outcomes for elderly and poor, chronically ill patients treated in HMO and fee-for-service systems: Results from the Medical Outcomes Study. JAMA 1996;276:1039–1047.


Key words: Health status; interpretation; calibration.