Health outcomes. New quality measure for Medicare

JAMES K. COOPER1,4, THOMAS KOHLMANN2, JAMES A. MICHAEL3, SAMUEL C. HAFFER1 AND MARCIA STEVIC1

1Health Care Financing Administration, Baltimore, MD, USA, 2University of Lübeck, Germany, 3Arizona Peer Review Organization (PRO), Phoenix, AZ, and 4the Uniformed Services University of Health Sciences, Bethesda, MD, USA

Abstract

Objective. A new measurement of health care quality for Medicare beneficiaries has been implemented by the Health Care Financing Administration (HCFA). This paper describes the program, presents baseline data and highlights associated issues.

Design. The Health Outcomes Survey (HOS) is a longitudinal cohort mail survey. Changes in population health status after 2 years will be evaluated on an individual plan level.

Setting. Two-hundred and eighty-seven US Medicare managed care plans.

Main outcomes measures. Physical component and mental component summary scales derived from the SF-36.

Findings. Baseline data documented lower health status in older populations, while functional limitations and disease prevalence were higher. Among different plans, mean functional levels were found to be similar, although a few plans contained populations with exceptionally low levels. These data do not support the assertion that enrollees in for-profit plans are healthier than non-profit plans.

Conclusions/implications. The HOS is the first large-scale program to evaluate health outcomes among older Americans. HCFA recognizes several technical and policy issues. Technical issues include possible biased reporting for subpopulations, the validity of proxy responses and respondent burden. Policy issues concern the appropriateness of using a generic measure such as the SF-36 and how much change in health status can be attributed to quality of health care. HCFA plans to extend the HOS to beneficiaries in traditional Medicare. The HOS project is expected to encourage more efforts to maintain or improve the health status of the Medicare managed care population.

Keywords: aged, health status, quality measurement, Medicare, SF-36

Health care for older Americans is largely financed by Medicare, which covers 34 million people aged 65 years and over. Medicare is administered by HCFA. HCFA has pursued health care quality improvement for Medicare beneficiaries for over 25 years. There remain, however, important opportunities for further quality improvements [1–3]. Evaluating the success of quality improvement programs depends on the availability of pertinent and acceptable quality measurement tools. Historically, both HCFA’s efforts, and evaluation of those efforts, have focused on health care structure and process [4]. Structure and process indicators do not measure quality of life and functional status, which are outcomes of high importance to both HCFA and Medicare beneficiaries.

To draw attention to outcomes as quality indicators, HCFA now requires Medicare managed care plans to obtain periodic outcomes measures from a sample of their enrollees. The new performance measure was originally called the Health of Seniors measure. Now, recognizing that some Medicare beneficiaries are under aged 65, it is called the Health Outcomes Survey (HOS), although the majority (approximately 87%) of Medicare beneficiaries are aged 65 years and over. The HOS is the first national measure of the quality of life and functional status of Medicare beneficiaries.

HCFA has three goals for this performance measure program: (i) to monitor health care quality in an ongoing standardized way; (ii) to provide information about quality to Medicare beneficiaries to assist them in choosing between different managed care plans; and (iii) to provide further incentive for quality improvement. HCFA believes that providing information about outcomes to beneficiaries will encourage them to select plans with better outcomes. This will produce market forces that favor plans with better results.

Address reprint requests to J. Cooper, Health Care Financing Administration, S3–02–01, 7500 Security Boulevard, Baltimore, MD 21244-1850, USA. E-mail: jcooper3@hcfa.gov

© 2001 International Society for Quality in Health Care and Oxford University Press
This will provide an incentive to plan designers to continually improve care and outcomes for their enrollees, especially older enrollees who make up the majority of Medicare beneficiaries. Aligning such financial incentives with health outcomes should improve outcomes [5].

The purpose of this report is to briefly describe the new health care performance measure, to present baseline data, and to discuss issues surrounding its use. The issues are relevant to most quality of care evaluations that incorporate generic health status measures. Because Medicare is mostly a program for people over the age of 65 years, the HOS is especially relevant to those interested in health care for older people.

Methods

Development and implementation of the HOS was accomplished in a public-private partnership with the National Committee for Quality Assurance (NCQA). NCQA is a not-for-profit managed care accreditation organization, and is perhaps best known for producing the annual Health Plan Employer Data and Information Set (HEDIS). HEDIS is a standardized set of performance measures that reflect the quality of health care provided in a health care plan [6,7]. NCQA’s HEDIS oversight body, the Committee on Performance Measurement, adopted and endorsed the HOS measure as a HEDIS 3.0 measure for Medicare plans.

The HOS collects health status information from individuals at one point in time and again 2 years later. The difference in self-reported health status is the ‘change score’. Adjustments in scores will be made for expected change, depending on baseline health conditions. For example, individuals with more health problems are expected to decline in health status at a rate greater than can be expected by age alone [8]. Details of this case-mix adjustment have not been finalized, but are expected to include demographics and disease variables. The impact of these variables on summary scores among the entire population will be the basis for the case-mix adjustment. On the basis of adjusted change scores, individual beneficiaries can be placed into one of three categories: improved, stayed the same, or declined in health status over the 2 years.

Performance scores for managed care plans will be calculated as the percentage of beneficiaries whose functional status improved, the percentage of beneficiaries whose functional status remained constant, and the percentage of beneficiaries whose functional status declined. These change scores for each health care plan in a market area will be made available to eligible beneficiaries to help beneficiaries make decisions when choosing a health plan.

The major element of the HOS measure is the SF-36, a 36 item survey questionnaire that has been used and validated in many studies [9–11]. The SF-36 appears to be the most widely used health status measure in the USA, and is used extensively worldwide [12]. It incorporates most of the widely used concepts of general health-related quality of life [13]. There are over 800 studies that have used the SF-36 [14]. It may be more sensitive to small differences in health status than other questionnaires [15] and, although it is a relatively short form, it appears to be as sensitive to small differences as longer, more time-consuming forms [16].

Responses to SF-36 questions are aggregated and combined with different weights to produce two summary scales, the Physical Component Summary scale and the Mental Component Summary scale [17]. These summary scales are believed to represent distinct clusters derived from the variance in responses to SF-36 questions. Reliability of these scales usually exceeds 0.90 [12]. The two Summary scales are used as the bases for change scores in the HOS project.

In addition to the SF-36 questionnaire, the HOS obtains data about an individual’s clinical case mix adjustment variables (complications, co-morbidities, chronic conditions); and demographic information. The SF-36 contains three Activities of Daily Living (ADL) questions; three additional ADL questions were added. The additional ADLs may be useful in determining fair reimbursement rates in the future, since individuals with more disability tend to use more resources. While the SF-36 is the core of the survey, the additional questions are meant to provide data for case-mix adjustment, and to increase understanding of determinants of health status as measured by the SF-36.

Survey design

The HOS measure is a longitudinal cohort survey administered to 1000 Medicare beneficiaries in each plan who have been continuously enrolled for at least 6 months. HCFA selects the sample and the survey is administered by independent survey research firms certified by NCQA. The sampling frame includes the aged and the disabled, but excludes those eligible for Medicare because of end stage renal disease. A questionnaire is mailed to selected participants. A reminder postcard and a second copy are mailed to non-respondents. Up to six telephone contacts are made for those still not responding. The survey is administered to the cohort at baseline and again 2 years later. A new cohort is selected each year for baseline and subsequent 2-year measurement. Wave one data collection occurred between 26 May–30 July 1998. Surveys were mailed to over 278 000 beneficiaries in 268 managed care plans covering 287 market areas nationwide; all managed care plans were surveyed.

Results

The response rate for the baseline 1998 cycle was 60% (range 29–79% among different plans). There were 167 069 individual responses from the 287 plans across the country. Only 1809 responses were from individuals in institutions, and only 63 of these were completed by the subject him/herself and were complete enough to calculate summary scores. Because of the small number, this institutionalized population is not described here. Of all respondents age 65 years and older, 85% were not living in institutions, and did
answer the questionnaire themselves. For this group, i.e. subjects age 65 years and older, not living in an institution, and who completed the survey themselves \( (n=141\ 589) \), the mean age was 73.7 years. All data reported here refer to this group.

For this group, the national mean Physical Component Summary Scale score was 42.6 (SD = 11.4). The national adult all-age norm is 50. Lower scores imply worse health. Among the 287 plans, mean Physical Component Summary Scale scores ranged from 33–45. One plan had an even lower mean score \( (28) \), but that plan is not representative because it had a low response rate due to administrative factors, and the plan’s purpose and enrollment are unique. The average of the mean plan scores was 41.2. While the range among plans was relatively wide, the distribution of mean plan Physical Component Summary scores was narrow. The SD for mean plan Physical Component Summary scores was 1.74. Kurtosis, a measure of ‘heaviness’ of the tails of the distribution, was 4.7, consistent with larger than usual ‘tails,’ and skewness was \(-1.2\), indicating the distribution had some skewness toward lower numbers. 5% of plans had a mean Physical Component Summary score of 39 or less. In summary, the distribution of mean plan Physical Component Summary scores was relatively normal except for some over-representation of plans with lower mean Physical Component Summary.

The average national Mental Component Summary score in this population was 53.3 (SD = 9.3). Among plans, mean Mental Component Summary Scale scores ranged from 44–55. The average of plan mean scores was 53 and (SD = 1.1). As with the Physical Component Summary Scale, the adult all-age norm is 50. Kurtosis was 1.165 and skewness was \(-1.19\), suggesting a more standard distribution of plan mean Mental Component Summary scores. These scores are computed using algorithms developed elsewhere from data for the general US adult population [18].

Physical Component Summary scores, Mental Component Summary scores, and other characteristics of this baseline cohort showed some variation across sex and age (Table 1). In this Table, ‘young’ refers to subjects who were aged 65–72 years and ‘old’ refers to subjects aged 73 years or greater. Older respondents had a lower mean Physical Component Summary score, while the Mental Component Summary Scale scores ranged from 33–45. One plan had an even lower mean score \( (31) \). 5% of plans had a mean Physical Component Summary score of 32 or less. In summary, the distribution of mean plan Physical Component Summary scores was narrow. The SD for mean plan Physical Component Summary scores was 0.4. Kurtosis, a measure of ‘heaviness’ of the tails of the distribution, was 1.74. Kurtosis, a measure of ‘heaviness’ of the tails of the distribution, was \(-1.19\), suggesting a more standard distribution of plan mean Physical Component Summary scores.

Comparing for-profit plans with not-for-profit plans \( (n = 103) \) revealed little difference in summary scores (Table 2). Enrollees in for-profit plans were very slightly more likely to have no limitations and no disease diagnoses, and to have less than high school education, but reported fair or poor health at the same rate as enrollees in not-for-profit plans. Profit and non-profit plans co-existed in five market areas. In four out of the five areas, profit plans had a lower mean Physical Component Summary score, indicating enrolled populations that were less healthy.

### Discussion

The HOS is a new, national health care performance measure for Medicare beneficiaries. The purpose of this report is to describe the measure, present baseline findings, and describe potential issues.

The lower than norm baseline mean Physical Component Summary Scale score of 42 for older Americans enrolled in Medicare managed care plans is not unexpected. Lower scores imply lower health status. As age increases, the score is expected to decrease [19]. In the Medical Outcomes Study, a study using 1986–1987 data, the baseline Physical Component scores for patients seen in different practice settings were 44.9 (prepaid system) and 45.2 (fee-for service) [8].

Baseline Physical Component mean scores among regular plans ranged from 33–45. The wide range documents that different managed care plans have different challenges in providing health care to older patients. It also emphasizes the need for efficient case-mix adjustment mechanisms for reimbursement, since responses to health status surveys strongly predict subsequent health care utilization [20].

It has been suggested that for-profit managed care plans might enroll more healthy people in order to achieve the best chances for profitability [21]. Our data do not support this suspicion as we found little difference in health status between profit and not-for-profit organizations.

Mental health among older Americans does not seem to deteriorate in parallel to physical health. The mean Mental Component Summary baseline score for the HOS (53) was above that of the USA norm (50).

The performance measures of the HOS are not the mean SF-36 scores of the plans. The performance measures will be the proportions of enrollees in each plan whose health status has improved, stayed the same, or declined after 2 years. The measure depends on individual differences between baseline and 2-year scores, or ‘change scores.’ Two-year scores will be adjusted based on age and medical conditions. Others have used this approach. The Medical Outcomes Study used a statistical approach to determine if follow-up scores changed more than would be expected by chance alone [22]. Using age, baseline functioning, self-rated health and co-morbidity, Mor and colleagues were able to predict functional status in elderly people 6 years later [23]. The ability to adjust expected health status on the basis of baseline functioning, self-rated health and co-morbidity is an important issue and is vital to the validity of the HOS as a quality of care performance measure. NCQA will convene an expert panel to make recommendations for defining the precise cut-off figures for each outcome category. The committee will use a combination of published data, statistical techniques, and expert judgment.

Several other issues have surfaced related to use of the HOS measure. These issues can be classified into two categories, technical and policy-related.

### Technical issues

Technical issues generally concern the possibility of inaccurate, non-valid, or unreliable scores from identifiable sub-populations such as the very frail, cognitively impaired [24].
**Table 1.** Baseline characteristics of respondents aged 65 years and older, who answered the questions themselves, and who did not live in an institution. Age is divided around the mean age for respondents, 73 years.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Young</th>
<th>Old</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Physical Component Summary Score (PCS)</td>
<td>42.2</td>
<td>44.2</td>
<td>40.2</td>
<td><em>P</em> &lt; 0.0001&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mean Mental Component Summary Score (MCS)</td>
<td>53.3</td>
<td>53.8</td>
<td>52.8</td>
<td><em>P</em> &lt; 0.0001&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mean number of ADL limitations (6 total)</td>
<td>0.78</td>
<td>0.61</td>
<td>0.94</td>
<td><em>P</em> &lt; 0.0001&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mean number of disease diagnoses (13 total)</td>
<td>2.45</td>
<td>2.27</td>
<td>2.62</td>
<td><em>P</em> &lt; 0.0001&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>Respondents reporting ADL limitations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero limitations</td>
<td>64.0%</td>
<td>70.6%</td>
<td>57.3%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>One limitation</td>
<td>14.8%</td>
<td>13.1%</td>
<td>16.6%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Two limitations</td>
<td>11.1%</td>
<td>8.9%</td>
<td>13.3%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Three or more limitations</td>
<td>10.1%</td>
<td>7.4%</td>
<td>12.8%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Respondents reporting disease diagnoses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero conditions</td>
<td>14.3%</td>
<td>16.5%</td>
<td>12.1%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>One condition</td>
<td>21.8%</td>
<td>23.6%</td>
<td>20.1%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Two conditions</td>
<td>21.3%</td>
<td>21.5%</td>
<td>21.0%</td>
<td>None</td>
</tr>
<tr>
<td>Three conditions</td>
<td>17.2%</td>
<td>16.1%</td>
<td>18.2%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Four conditions</td>
<td>11.4%</td>
<td>10.3%</td>
<td>12.5%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Five conditions</td>
<td>6.9%</td>
<td>6.0%</td>
<td>7.9%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Six conditions</td>
<td>3.7%</td>
<td>3.0%</td>
<td>4.4%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Seven or more conditions</td>
<td>3.4%</td>
<td>3.0%</td>
<td>3.8%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>High school education</td>
<td>73.6%</td>
<td>76.9%</td>
<td>70.3%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Health is ‘fair’ or ‘poor’</td>
<td>23.4%</td>
<td>19.0%</td>
<td>27.8%</td>
<td><em>P</em> &lt; 0.001&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>1</sup>T-test of differences in means between young and old.

<sup>2</sup>Chi square test of differences in proportions between young and old.

or minority culture populations. While adjustments for race and ethnicity can be made, such adjustments may not fully compensate for cultural and other subpopulation characteristics. Further, adjusting for race and ethnicity may hide important differences that deserve consideration. If an effect, good or bad, is associated with an explicit population segment, adjusting for the population segment will conceal the effect. This could mask discrimination. It is important to examine summary data using stratified or multivariate analyses to examine possible racial or ethnic associations.

### Disadvantaged and minority populations

Research on use of the SF-36 in disadvantaged populations is limited. Use of the SF-36 in a disadvantaged population attending a general medicine outpatient clinic in an academic medical center was reported recently. Results indicated that the instrument was sensitive enough to measure change, but little of the change was predictable [25]. Baseline scores for this group were well below national norms. Investigators found internal consistency to be unacceptable for the general health perceptions scale, but were satisfied that the SF-36 generally was suitable for measuring individual change in disadvantaged older adults with debilitating chronic disease. Whether these findings are generalizable remains to be determined.

The SF-36 has not been widely used in all American minority populations. While the SF-36 has been found to be reliable across Western European countries [26], it does not produce comparable responses everywhere. For example, among older Japanese men, the SF-36 may fail to accurately assign problems in role function to mental conditions [27]. Investigators have suggested this may be due to culture-based difficulty accepting that emotional problems can cause functional limitations. The possibility of culture-based response differences in American subcultures has not been excluded.

### Proxy respondents

Another issue is reporting source bias. It is likely that responses from the most frail patients will actually come from family or care-givers acting for the patient. Family members have been seen to view impairments or disabilities differently than patients themselves. Often they see impairment as more severe [28,29], although some analyses indicate proxy responses do not affect accuracy [30]. The effect on change scores is not known.

### Respondent burden and response rate

While most respondents can complete the survey in 20 minutes, respondents with physical or emotional disabilities may require longer. Many suspect that the accumulated burden of completing long surveys may cause a differential response rate for individuals with higher levels of disabilities – they
Table 2. Comparison between for profit and not for profit Medicare managed care plans. Variables are summary scores, ADL, number of chronic conditions, education and self-report of health [‘In general, would you say your health is: (select) excellent, very good, good, fair, poor’]

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>For profit</th>
<th>Not for profit</th>
<th>Significant differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Physical Component Summary Score (PCS)</td>
<td>42.2</td>
<td>42.3</td>
<td>42.1</td>
<td>No</td>
</tr>
<tr>
<td>Mean Mental Component Summary Score (MCS)</td>
<td>53.3</td>
<td>53.3</td>
<td>53.3</td>
<td>No</td>
</tr>
<tr>
<td>Mean number of ADL limitations (6 total)</td>
<td>0.78</td>
<td>0.77</td>
<td>0.79</td>
<td>No</td>
</tr>
<tr>
<td>Mean number of disease diagnoses (13 total)</td>
<td>2.45</td>
<td>2.44</td>
<td>2.46</td>
<td>No</td>
</tr>
</tbody>
</table>

Respondents reporting ADL limitations
- Zero limitations: 64.0% (for-profit), 64.3% (not-for-profit), 63.3% (All), *P* < 0.001
- One limitation: 14.8% (for-profit), 14.6% (not-for-profit), 15.3% (All), *P* < 0.001
- Two limitations: 11.1% (for-profit), 10.9% (not-for-profit), 11.3% (All), *P* < 0.05
- Three or more limitations: 10.1% (for-profit), 10.2% (not-for-profit), 10.1% (All), No

Respondents reporting disease diagnoses
- Zero conditions: 14.3% (for-profit), 14.5% (not-for-profit), 14.0% (All), *P* < 0.05
- One condition: 21.8% (for-profit), 21.9% (not-for-profit), 21.7% (All), No
- Two conditions: 21.3% (for-profit), 21.2% (not-for-profit), 21.3% (All), No
- Three conditions: 17.2% (for-profit), 17.1% (not-for-profit), 17.3% (All), No
- Four conditions: 11.4% (for-profit), 11.3% (not-for-profit), 11.5% (All), No
- Five conditions: 6.9% (for-profit), 6.8% (not-for-profit), 7.1% (All), No
- Six conditions: 3.7% (for-profit), 3.7% (not-for-profit), 3.8% (All), No
- Seven or more conditions: 3.4% (for-profit), 3.5% (not-for-profit), 3.3% (All), No
- High school education: 73.6% (for-profit), 72.3% (not-for-profit), 75.8% (All), *P* < 0.001
- Health is ‘fair’ or ‘poor’: 23.4% (for-profit), 23.6% (not-for-profit), 23.2% (All), No

1 T-test of differences in means between for-profit MCOs and not-for-profit MCOs.
2 Chi square test of differences in proportions between for-profit MCOs and not-for-profit MCOs.

may be less likely to complete the survey. Our data do not show a response rate bias in summary scores among plans – plans with higher response rates did not have consistently lower summary scores. From another perspective, the mean Physical Component Summary scores of all subjects who responded to the first mailing was 41.6 (95% CI, 41.5–41.7); for those who responded to the second mailing it was 40.1 (95% CI, 40.5–40.8). Scores from the second mailing were significantly lower than the first (*P* < 0.0001). The mean score from the telephone inquiry round was 41.2 (95% CI, 40.9–41.3), and to the express mail round was 40.2 (95% CI, 39.5–40.8). Since all rounds after the first round had lower mean Physical Component Summary scores, we have interpreted these results as suggesting there is some response rate bias. We do not know the health status of those who failed to respond at all (non-responders), but we know they are older (mean age = 75.1 years), have a higher proportion of African-Americans (11.2% compared with 7.5%), and are more likely to be on Medicaid, a social insurance program predominantly for low income individuals (4.3% compared with 2.9%). All these differences are significant (*P* < 0.001).

There is no difference in sex between responders and non-responders. This raises two important issues: (i) how much do response rate and non-response rate bias results and (ii) if results have been biased, would shorter questionnaires overcome the problem? We do not know if the full SF-36 questionnaire is required for Plan comparisons, or if a shorter version, such as the SF-12, would be sufficient. This issue is especially important since individual change scores require that the respondent complete both the baseline and the 2 year follow-up survey. If more ill or otherwise unique individuals are more likely to ‘drop out’, the performance measure will be less reliable.

**Policy issues**

The first community health indicators were ‘vital statistics’ that included mortality rates and were used principally by health departments and government agencies. Improvements in mortality rates in this century have been largely due to
improvements in sanitation, pre-natal care and the use of antibiotics [31]. Mortality rates do not reflect differences in modern clinical care for chronic diseases and therefore have not been especially useful for evaluating differences in clinical care for older people. But what is the best choice is for an indicator to evaluate health care of older people is still being discussed.

**Disease specific or generic measures?**

The SF-36 is a generic measure. If the goal is to improve health care quality, disease specific measures, rather than generic measures, might be more useful. Health care in America is largely disease specific; students are taught about diseases, clinicians diagnose and treat diseases, treatment guidelines are disease-oriented. Disease specific measures determine how well clinicians do. For example, arthritis treatment can be evaluated by dimensions such as walking speed and grip strength. Disease specific measures are more likely to detect smaller changes than generic measures, and are likely to be more sensitive to different treatment approaches [32]. For evaluation of specific clinical care of chronic diseases, disease specific measurements such as range of motion for arthritis treatment and glycosylated hemoglobin for diabetes treatment are more often used than generic outcomes measures.

But while useful in some instances, disease specific measures represent only a segment of a person’s life, and are not relevant to people without the specific disease. For older populations, there has been increasing interest in measuring the impact of disease on functional capacity and quality of life [33]. For these reasons, generic outcomes measures have been sought. Arguing against generic measures, some say the questionnaires used do not measure function as much as they measure willingness to complain about perceived functional disability [34]. So far, it has been unusual for generic outcomes to be measured in clinical trials [35]. But evaluation of clinical health care has evolved, and outcomes measures are beginning to take their place as legitimate measures of the end results of health care [36].

**Is change score the best measure?**

Using individuals as their own ‘control,’ i.e. comparing change scores rather than current scores, is appealing because it would seem inherently to adjust for individual baseline variation. It accomplishes individual risk adjustment. But use of change scores may not be the best measure of treatment effect. If there is more error variation than variance between patients, using change scores may introduce even more error in estimating effects of treatment [37].

**Should nursing home residents be included?**

Another issue is whether to include those individuals who reside in long term care facilities when calculating the public quality measure. Data in the tables presented here do not include institutionalized subjects. Individuals in long term care generally have a lower Physical Component Summary score. The mean Physical Component Summary score of institutionalized people in the baseline survey was 34. Long term care residents have a shorter life expectancy. Currently, death produces an individual change score of ‘worse’. Including long term care residents is likely to increase the proportion of people who ‘got worse’ in a plan. Including them in this quality monitoring and health care performance measure may be unfair to plans that enroll more institutionalized people, and may lead to disincentives to their enrollment. Not including them in this quality monitoring and health care performance measure may deflect plan attention from them, and leave them out of quality improvement efforts.

**Attribution**

Do instruments such as the SF-36 measure an outcome that can be directly attributed to actions of a health care plan? Brook and colleagues argue that the best assessments of care quality are based largely on process criteria, not outcomes criteria [38]. A small study (n = 201) of disability prevention and disease self-management in frail older people found a significant reduction in inpatient hospital days but not a significant difference in SF-36 scores [39]. A recent study suggests that with aging, as the relative importance of genetic factors diminishes, environmental, health care and lifestyle factors become dominant [40]. For example, socioeconomic status has been shown to affect SF-36 scores.

In one study, men without known disease scored lower in the physical functioning scale if they were in the lowest employment grades [41]. Not having access to a car was also linked to low scores even after adjusting for employment grade. This suggests that what is measured by the SF-36 is not completely explained by infirmity or disease, and is, to a greater or less extent, determined by individual circumstances, i.e. by factors outside the body’s physiology and pathology. Thus, what is measured by the SF-36 does not ‘stop at the skin,’ as had been hoped [13]. Is it fair for health plans to be held accountable for individual health status, since health care is only one of the determinants of health status, and the other determinants are out of the control of health plans? Clinicians may assert that it is not reasonable to expect clinical providers to maintain the health status of those who choose to smoke, to not exercise, and to have a poor diet, for example. On the other hand, factors ‘outside the skin’ may be largely adjusted-for by considering only the change in score, since external factors are likely to affect baseline and 2 years scores equally.

Many now assert that, at least for chronic diseases, outcomes research and measurement will change clinical practice in the USA [42]. Generic outcomes measures are increasingly being used to evaluate health care. Many clinical trials have demonstrated drug or procedure treatment responses in small populations using the SF-36. Systematic interventions such
as routine reminder systems and patient education have improved outcomes in populations of chronically ill people [43]. However, so far there are not many examples to demonstrate that generic outcomes in older people are sensitive to health care organization and delivery [33]. The HOS project itself is a test of this issue. If there is variation in distribution of change scores among plans as expected, i.e. if some plans have significantly fewer members who ‘got worse,’ Medicare-supported Professional Review Organizations (PROs) and others will attempt to characterize features of those plans that contributed to their success. As other plans replicate those features, the effectiveness of the features will be evaluated. HCFA plans to extend the HSO to include beneficiaries who are in traditional Medicare as well as managed care. In the meantime, clinicians whose care may be judged by SF-36 change scores understandably may be suspicious that the link between SF-36 change scores and health care has not been sufficiently established. HCFA believes the link has been established, and health care plans will respond by focusing on ways to improve general health status. This issue remains in the realm of debate because, at this time, conclusions about the assertion have not been universally accepted.

**Conclusion**

The HOS performance measure is based on the premise that an adequate linkage exists between health care and health status of Medicare beneficiaries, or at least between health care and change in health status as measured by generic measures such as the SF-36 questionnaire. Technical issues have been described for the new HOS. Illuminating these issues with discussion and research will increase our understanding of the utility of these measures. Policy issues are also apparent, as in any national program. The HOS uses a large database for quality monitoring. As Davidoff points out, ‘... the use of large databases for medical quality improvement is a true innovation, the future of which will probably be determined as much by the social and emotional forces that govern the diffusion of all innovations’. [44]. HCFA relies on input from experts in many fields to reach the best decisions about these issues, and how to measure and improve care generally. But perhaps most importantly, it may be that the discussion that surrounds the HOS performance measure will lead to innovations that clearly demonstrate ways to prevent or delay suffering, care burden, and costs from chronic disease in older people. Simply choosing an indicator tends to affect the health of the population, since the indicator focuses attention to that problem and encourages new interventions [45].

**References**

19. McHorney CA, Kosinski M, Ware JE. Comparisons of the costs and quality of norms for the SF-36 health survey collected by


Accepted for publication 27 September 2000