Prospective versus retrospective measurement of change in health status: a community based study in Geneva, Switzerland

Thomas V Perneger, Jean-François Etter, André Rougemont

Abstract

Study objectives—To compare prospective and retrospective measurements of change in health status.

Design—Health status was measured using a French language version of the short form 36 (SF-36) health survey on two occasions one year apart—in 1992 and 1993. Differences in SF-36 scores measured prospectively were compared with the patients’ single item retrospective evaluation of change in health (transition item).

Setting—This was a community based study among members of two health insurance plans in Geneva, Switzerland.

Participants—Altogether 831 young adults (mean age 30 years at baseline).

Main results—Health status remained stable on average during the study period. The prospective rating correlated well with changes in health measured prospectively: those who said in 1993 that their current health was “much worse” than in 1992 experienced an average decrease of 1.06 SD on the eight SF-36 scales, while those who said that their health was “much better” recorded an average improvement of 0.43 SD. The associations between prospective and retrospective assessments of change were approximately linear for all scales but physical functioning. The transition item also discriminated between time periods: transition reported for 1991–92 did not correlate with changes recorded for 1992–93. Relative validity analyses indicated that the transition item was better suited to capture changes in general health than changes in purely physical or mental aspects of health.

Conclusions—The concordance between prospective and retrospective measures of change in health suggests that both are sensitive, to some extent, to true changes in health status. Using both types of assessment may improve the reliability of measurements of change.

(J Epidemiol Community Health 1997;51:320–325)

Generic health questionnaires are potentially valuable tools for assessing the effectiveness of health services in the community. The SF-36 health survey (SF-36)1–3 has been recommended for such use.4 To provide useful feedback, an indicator must be able to detect change over time,5–8 but until now, the sensitivity to change of the SF-36 has been examined only in studies of treatments likely to induce large improvements in physical functioning, such as total knee9 or hip10,11 replacement. The responsiveness of the SF-36 in community based settings is unknown.

The SF-36 includes 35 items that measure eight dimensions of health and one retrospective transition item: “Compared to one year ago, how would you rate your health in general now?” Possible answers are: “Much better now than one year ago,” “Somewhat better . . .,” “About the same . . .,” “Somewhat worse . . .,” and “Much worse . . .” The validity of the health transition item has not been closely examined.12

In the course of a prospective study comparing two health insurance plans, we administered a French language version of SF-36 on two occasions, one year apart, to a cohort of young adults. In this paper, we describe changes in SF-36 health dimension scores observed in a fairly healthy population, examine the properties of the retrospective transition item, and compare prospective and retrospective assessments of change in health.

Methods

STUDY SETTING AND DATA COLLECTION
This analysis was conducted as part of a controlled prospective study of the impact of a new, university sponsored managed care organisation on the use of health services, health status, and satisfaction with care.13 Randomly selected members (n = 1235) of the managed care plan and of the control indemnity insurance plan who were French speaking residents of Geneva, Switzerland and were aged 18–44 years were sent a mailed survey in November 1992, and 1007 (82%) participated. In November 1993, 959 (95%) of the 1992 respondents still lived in Geneva and were contacted for a follow up survey. A total of 851 (89%) of them responded for a second time. Twenty persons were eliminated because of a discrepancy in age or sex; the remaining 831 respondents (87% of eligible persons) constitute the sample used in this report.

Respondents and non-respondents to the follow up survey were similar in terms of age, sex, and baseline health status. The mean (SD) age of the respondents was 30 (5.7) years at the time of the baseline survey, 444 (53%) were women, 438 (53%) were university students, and 609 (73%) worked either part time or full time.
Table 1  Scores for eight dimensions of health measured one year apart using the MOS 36-item short form health survey (French version) in 831 young adults, Geneva, Switzerland, 1992 and 1993

<table>
<thead>
<tr>
<th>No. with both measurements</th>
<th>Mean (SD) score in 92</th>
<th>Mean (SD) score in 93</th>
<th>Intraclass correlation coefficient</th>
<th>Internal consistency coefficient*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical functioning</td>
<td>819</td>
<td>91.5 (17.5)</td>
<td>90.9 (17.8)</td>
<td>-0.6 (16.7)</td>
</tr>
<tr>
<td>Role—physical</td>
<td>812</td>
<td>87.0 (27.1)</td>
<td>85.8 (29.5)</td>
<td>-1.2 (33.7)</td>
</tr>
<tr>
<td>Bodily pain</td>
<td>822</td>
<td>78.8 (22.9)</td>
<td>77.6 (22.8)</td>
<td>-1.3 (23.9)</td>
</tr>
<tr>
<td>General health</td>
<td>799</td>
<td>77.0 (17.8)</td>
<td>76.1 (18.9)</td>
<td>0.0 (13.9)</td>
</tr>
<tr>
<td>Vitality</td>
<td>814</td>
<td>65.2 (17.2)</td>
<td>65.1 (17.4)</td>
<td>0.0 (16.2)</td>
</tr>
<tr>
<td>Social functioning</td>
<td>824</td>
<td>84.1 (20.4)</td>
<td>83.7 (21.3)</td>
<td>-0.5 (22.6)</td>
</tr>
<tr>
<td>Role—emotional</td>
<td>812</td>
<td>77.0 (34.2)</td>
<td>79.2 (34.2)</td>
<td>-2.2 (38.2)</td>
</tr>
<tr>
<td>Mental health</td>
<td>815</td>
<td>69.3 (18.3)</td>
<td>69.2 (17.9)</td>
<td>0.0 (16.2)</td>
</tr>
</tbody>
</table>

* Cronbach α, from 1992 survey

HEALTH STATUS MEASUREMENT INSTRUMENT

A translated version of the SF-36 was used to measure eight dimensions of health. Each scale score lies between 0 (worst possible health) and 100 (best possible health). One item refers to perceived change in health over the past year. The translation into French and the properties of the translated instrument have been reported on elsewhere. Completion rates, reliability coefficients, and tests of validity were similar to those reported for the original American version. The wording of several items of this version differs from the French translation produced by the international quality of life assessment project, which was not available when this study was conducted.

ASSESSMENT OF CHANGE IN EIGHT DIMENSIONS OF HEALTH

For each dimension of health status, we examined the distributions of score differences between 1992 and 1993, and computed the intraclass correlation coefficient for each scale using ANOVA models. This "test–retest" correlation with a one year interval between measurements reflects both test–retest reliability and actual changes in health status over the period of observation. This analysis was repeated in a subset of participants who reported that their health had not changed during the year of study—i.e. the most stable group. Scale reliabilities were estimated by Cronbach α coefficients.

PROPERTIES OF HEALTH TRANSITION RATINGS

We examined the distribution of responses to the transition item, hypothesising that in a steady state population, the proportion of respondents stating that their health status is "better" and "worse" than last year should be about the same. A skewed distribution of responses would indicate either bias in health transition ratings or lack of a steady state.

Then, we examined the association between health transition and current health (general health scale). We hypothesised that self perceived improvement in health should be moderately correlated with higher values of current health status, but that each rating of change ("same", "better", "worse") should be observed over a wide range of current health status scores. Box plots were used to examine the distributions of current health scores.

Concordance between prospective and retrospective measures of change

The 1992–93 differences in the eight health status scores were examined across retrospective health transition ratings. The sensitivity of each scale to change defined by each transition rating was assessed using the responsiveness index, given by the ratio of the average change to the standard deviation (SD) of difference scores observed in stable subjects (those who reported their current health to be "about the same" as one year ago). To assess the ability of retrospective ratings to discriminate between specific time periods, we compared the 1992–93 difference for general health to the transition rating given in 1993 (in this case, we expected a positive correlation) and to the transition rating given in 1992 (in that case, we expected little or no correlation).

Using means of the eight scale-specific responsiveness indices as the reference, we tested the hypothesis that transition ratings were on an interval scale—i.e., that the difference between "much worse" and "somewhat worse" is the same as the difference between "somewhat worse" and "about the same". Interval scaling is a prerequisite for all basic arithmetic operations.

Last, we assessed the relative validity of the eight prospective measurements of change in explaining retrospective ratings of health status change. For each scale, a simple linear regression model of retrospective ratings (coded 1 through 5) on prospectively measured differences was computed. The variation in retrospective ratings explained by each scale was estimated by the squared r statistic from each linear slope estimate. Relative validity was expressed as a ratio of variance explained by a
Table 2  Average change in eight dimensions of health measured one year apart using the short form 36 health survey (French version) in 831 young adults, for free retrospective health transition ratings, Geneva, Switzerland, 1992 and 1993

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Much worse</th>
<th>Somewhat worse</th>
<th>About the same</th>
<th>Somewhat better</th>
<th>Much better</th>
<th>p value*</th>
<th>Between group differences</th>
<th>Linear trend</th>
<th>Deviation from linearity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical functioning</td>
<td>-16.2**</td>
<td>-4.6*</td>
<td>0.1</td>
<td>-0.2</td>
<td>-0.9</td>
<td>0.02</td>
<td>-0.01</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Role—physical</td>
<td>-35.7**</td>
<td>-11.8*</td>
<td>-1.6</td>
<td>0.1</td>
<td>15.5**</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Bodily pain</td>
<td>-20.5*</td>
<td>-9.9***</td>
<td>-1.4</td>
<td>-0.0</td>
<td>8.2**</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>0.25</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>General health</td>
<td>-21.3**</td>
<td>-10.1***</td>
<td>-0.8</td>
<td>0.6</td>
<td>8.1**</td>
<td>0.01</td>
<td>&lt;0.001</td>
<td>0.23</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Vitality</td>
<td>-14.2*</td>
<td>-5.0*</td>
<td>-1.0</td>
<td>1.7</td>
<td>11.1**</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.42</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Social functioning</td>
<td>-23.2*</td>
<td>-7.0</td>
<td>-1.5</td>
<td>3.8</td>
<td>8.2**</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.01</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Role—emotional</td>
<td>-33.3*</td>
<td>-0.0</td>
<td>0.4</td>
<td>9.5*</td>
<td>8.3</td>
<td>0.009</td>
<td>0.004</td>
<td>0.17</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Mental health</td>
<td>-8.4</td>
<td>-2.7</td>
<td>-0.9</td>
<td>1.5</td>
<td>8.2***</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>0.48</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Difference with middle category ("about the same"): *p < 0.05; **p < 0.01; ***p < 0.001.
†From one way ANOVA.

Results

PROSPECTIVE CHANGE IN EIGHT DIMENSIONS OF HEALTH STATUS

Most of the 831 respondents provided sufficient data (ie, a half or more of items for each scale) to compute baseline (1992) and follow up (1993) scores for the eight dimensions of health (table 1); 775 (93%) participants had no missing values. Mean scores were similar in 1992 and 1993; differences ranged from -0.8 to +2.2 on a 0–100 scale, and none was statistically significant. The intraclass correlation coefficient between measurements made one year apart was highest (0.70) for general health and lowest (0.28) for role—physical. For each scale, the intraclass correlation coefficient (test–retest) was lower than the corresponding Cronbach α coefficient, which estimates scale reliability from the internal consistency of its items. Among those who stated in 1993 that their health status was about the same as one year earlier, health status remained stable, just as in the whole sample, and intraclass correlation coefficients were similar (data not shown).

RETROSPECTIVE TRANSITION RATING

In 1993, 64 participants (7.7%) rated their current health as “much better”, 130 (15.6%) as “somewhat better”, 549 (66.1%) as “the same as”, 70 (8.4%) as “somewhat worse”, and 8 (1.0%) as “much worse” than one year ago. Ten (1.2%) did not respond. Women gave slightly higher transition ratings (mean 3.26 on scale from 1 to 5) than men (mean 3.15, p = 0.02). Transition ratings did not differ by age group, income category, or employment status (working versus student).
Table 3: Responsiveness index* of the eight dimensions of health measured one year apart using the short form 36 health survey (French version) in 831 young adults, as a more retrospective health transition ratings, Geneva, Switzerland, 1992 and 1993

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Much worse</th>
<th>Somewhat worse</th>
<th>About the same</th>
<th>Somewhat better</th>
<th>Much better</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical functioning (RF)</td>
<td>-1.01</td>
<td>-0.28</td>
<td>0.00</td>
<td>-0.01</td>
<td>-0.05</td>
</tr>
<tr>
<td>Role—physical (RP)</td>
<td>-1.25</td>
<td>-0.41</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.53</td>
</tr>
<tr>
<td>Bodily pain (BP)</td>
<td>-0.86</td>
<td>-0.42</td>
<td>-0.06</td>
<td>0.00</td>
<td>0.35</td>
</tr>
<tr>
<td>General health (GH)</td>
<td>-1.71</td>
<td>-0.79</td>
<td>-0.06</td>
<td>0.05</td>
<td>0.63</td>
</tr>
<tr>
<td>Vitality (VT)</td>
<td>-0.98</td>
<td>-0.35</td>
<td>-0.07</td>
<td>0.12</td>
<td>0.77</td>
</tr>
<tr>
<td>Social functioning (SF)</td>
<td>-1.13</td>
<td>-0.34</td>
<td>-0.07</td>
<td>0.18</td>
<td>0.40</td>
</tr>
<tr>
<td>Role—emotional (RE)</td>
<td>-0.96</td>
<td>-0.00</td>
<td>0.01</td>
<td>0.27</td>
<td>0.24</td>
</tr>
<tr>
<td>Mental health (MH)</td>
<td>-0.55</td>
<td>-0.18</td>
<td>-0.06</td>
<td>0.10</td>
<td>0.54</td>
</tr>
<tr>
<td>Average responsiveness†</td>
<td>-1.06</td>
<td>-0.35</td>
<td>-0.04</td>
<td>0.09</td>
<td>0.43</td>
</tr>
<tr>
<td>Empirical score‡</td>
<td>1.0</td>
<td>2.9</td>
<td>3.7</td>
<td>4.1</td>
<td>5.0</td>
</tr>
</tbody>
</table>

* Ratio of average change in group to the between-subject variability of difference scores in the group that rated current health to be “about the same” (SDs were: PF: 16.1, RP: 28.7, BP: 23.5, GH: 12.8, VT: 14.5, SF: 20.6, RE: 34.7, MH: 15.4).
† Simple mean of eight independent indices of responsiveness.
‡ Lowest (1.0) and highest (5.0) scores arbitrary; other scores proportional to average responsiveness.

Current general health scores were progressively higher in persons who rated their current health to be “much worse” (mean 31.6), “somewhat worse” (56.9), and “the same” (79.0) as a year ago. Respondents who rated their health as “somewhat better” had lower general health scores (73.7) than those who reported no change (p = 0.004 for a comparison of these two groups), while those who reported “much better” health had the highest scores (81.4). The range of general health scores was wide for each transition rating (fig 1).

PROSPECTIVE VERSUS RETROSPECTIVE MEASUREMENT OF HEALTH TRANSITION

Prospectively measured changes were associated with retrospective ratings for seven of the eight scales (table 2). The exception was physical functioning, for which the association appeared to be non-linear and was less statistically significant. Prospectively measured changes were generally similar in persons who rated their current health as “about the same” and “somewhat better”; nevertheless, deviations from linearity were not statistically significant, except for physical functioning and general health.

Responsiveness indices (crude change scores divided by the SD of change scores in persons who reported current health to be “about the same” as last year) were ordered in a logical fashion (table 3). On average, a transition rating of “much worse” corresponded to a responsiveness index of −1 SD, and “much better” corresponded to a responsiveness index of +0.5 SD. The range of responsiveness indices between extreme transition ratings was largest for general health, and narrowest for purely physical or mental dimensions of health. Empirically, using means of responsiveness indices as the reference, the optimal scoring of the five transition ratings would be 1.0, 2.9, 3.7, 4.1, and 5.0. Thus the assumption of interval scaling did not exactly fit the data. In particular, the interval between “about the same” and “somewhat better” was narrower than expected.

Respondents discriminated correctly between time periods (fig 2). The change in general health, measured prospectively between 1992 and 1993, was strongly associated with retrospective transition ratings covering that period (linear trend, p<0.001), but not with transition ratings covering the previous year (p = 0.13).

The relative validity of prospective measurements in explaining variation in retrospective transition ratings was examined using linear regression models (table 4). Changes in general health were best at predicting retrospective ratings, while changes in physical functioning were worst. In general, scales which measured purely physical or mental aspects of health explained less variation in retrospective ratings than did scales reflecting global aspects of health, such as general health or vitality. Relative validity results from polychotomous logistic regression models were similar (not shown).

Discussion

This study compared prospective and retrospective measurements of change in health obtained using the SF-36 in a general population. Results indicate substantial agreement between
these two approaches to the measurement of change. The agreement was not perfect, because neither measure is perfectly reliable, and because the retrospective assessment depends on the respondent's recall. Nonetheless, respondents' recall was generally consistent with changes in health status measured prospectively. Conversely, measurement of SF-36 scores one year apart was able to discern average differences in health that were perceptible to respondents after the fact. Thus both measures were responsive to some extent, which is an important aspect of an instrument's validity.23

Health status scores remained unchanged on average during follow up, but individual measurements taken one year apart were only moderately correlated. Test–retest differences arise from both random measurement errors and true differences in health status. Therefore a low test–retest correlation may be interpreted as evidence of low reliability, high responsiveness (especially when the time elapsed between measurements is long), or both. The contribution of responsiveness can be indirectly deduced from the comparison of internal consistency coefficients and test–retest coefficients. The two estimates of reliability should be equal if the instrument is insensitive to change23; the fact that test–retest coefficients were lower suggests that differences in SF-36 scores captured some true changes in health status.

Responses to the transition item were positively skewed; more people reported an improvement in health than a deterioration, despite stable health status in the whole cohort. A similar finding has been reported by others.12 The most likely explanation is response bias, attributable either to the social desirability of doing better or to the response format, which may favour the first response options. The transition rating was also related to current health status: those who rated their current health as "much worse" than last year had considerably lower general health scores than persons who rated their current health as "much better". The association was not linear: the high general health scores in persons whose health remained "about the same" can probably be attributed to the absence of significant health problems in most of our participants.

The central finding of this study is that transition ratings correlated about as expected with differences in prospective measurements. This correlation was specific for the time period under consideration. Using the transition item as the criterion of change, responsiveness indices were larger for general health than for any other SF-36 scale. Changes in general health had also the largest relative ability to explain variability in the transition item. This result is logical, because the transition item is worded in general terms, without referring to any specific aspect of health. In contrast, changes in physical functioning correlated rather poorly with health transition, perhaps due to the generally excellent physical health of our participants. If change in specific aspects of health is a concern, it may prove useful to supplement the SF-36 with dimension-specific transition items such as, "Compared with your physical health one year ago, how would you rate your physical health now?" Such transition ratings have been used previously.25 Interestingly, our result suggests that when general health status in an unselected population is of concern, generic health status instruments may be more responsive than more specific scales. Conversely, in patients with selected health problems such as benign prostatic hyperplasia,26 lumbar spinal stenosis,27 or cataracts,28 condition-specific instruments are the more responsive to change.

The five response modalities of the transition item were ordered on an ordinal scale, but the intervals between responses were unequal. Intervals, as reflected by prospective assessments, were greater for negative than for positive transitions, and transition ratings of "about the same" and "somewhat better" were both consistent with no change in health. The latter finding may reflect the social desirability of "doing better" even when no substantial change in health has occurred. For instance, patients tend to be optimistic in retrospective assessments of health status improvement after surgery.29 In our study, SF-36 scales were better at detecting self reported deteriorations in health than self reported improvements. A "ceiling phenomenon" may have contributed to this result: when baseline health status is very good, further improvements may be difficult to detect. A greater ability to detect declines in health was also reported in two other studies which used other instruments.12,30 If an instrument's responsiveness depends on the direction of change, collapsing positive and negative changes when assessing responsiveness31 may be unwise.

In this analysis, we used one of several available measures of sensitivity to change.32 All are standardised score differences, which vary in the manner of standardisation. The "effect size" is the crude difference between baseline and follow up scores divided by the SD of baseline scores.33 This measure reflects the composition of the population under study: an instrument will appear to be more responsive in a homogeneous population than in a population in which health states differ at baseline. The "standardised response mean" is the crude difference divided by the standard deviation of differences between baseline and follow up scores.34 It depends on the heterogeneity of changes in health: an instrument will appear more responsive when everybody changes in the same way than when some individuals get worse and others get better. The "responsiveness index" is the crude difference divided by the SD of differences between baseline and follow up scores in stable individuals, identified by independent means.6,19 This measure, which we used in this study, is most independent of population characteristics, but it requires an independent means of identifying stable individuals. This leads to a somewhat circular argument, since if we could assess stability in a valid and reliable way, we would not need another instrument to measure change. We used the transition item to select the "stable" group, but because a single item measure is necessarily unreliable, our "stable"
Measurement of health status change

Funding: Supported by the Swiss National Science Foundation (grants 3233-32609.91 and 32-39092.93).

5 Garrett AM, Ruta DA, Abdalla MI, Buckingham JK, Russell IT. The SF36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? BMJ 1993;306:1440-44.
10 Lanksy D, Butler JB, Waller FT. Using health status measures in the hospital setting: from accuracy to outcome and comes management. Med Care 1992;30(Suppl.):MS57-73.
33 Kanfer HI, Aminoff JJ, Moanen RF. Effect sizes for in- terpreting changes in health status. Med Care 1989;27(Suppl.):i174-89.