SF 36 health survey questionnaire: I. Reliability in two patient based studies

Danny A Ruta, Mona I Abdalla, Andrew M Garratt, Angela Coutts, Ian T Russell

Abstract

Objective—To assess the reliability of the SF 36 health survey questionnaire in two patient populations.

Design—Postal questionnaire followed up, if necessary, by two reminders at two week intervals. Retest questionnaires were administered postally at two weeks in the first study and at one week in the second study.

Setting—Outpatient clinics and four training general practices in Grampian region in the north east of Scotland (study 1); a gastroenterology outpatient clinic in Aberdeen Royal Hospitals Trust (study 2).

Patients—1787 patients presenting with one of four conditions: low back pain, menorrhagia, suspected peptic ulcer, and varicose veins and identified between March and June 1991 (study 1) and 573 patients attending a gastroenterology clinic in April 1993.

Main measures—Assessment of internal consistency reliability with Cronbach’s α coefficient and of test-retest reliability with the Pearson correlation coefficient and confidence interval analysis.

Results—In study 1, 1317 of 1746 (75·4%) correctly identified patients entered the study and in study 2, 549 of 573 (95·8%). Both methods of assessing reliability produced similar results for most of the SF 36 scales. The most conservative estimates of reliability gave 95% confidence intervals for an individual patient’s score difference ranging from −19 to 19 for the scales measuring physical functioning and general health perceptions, to −65·7 to 65·7 for the scale measuring role limitations attributable to emotional problems. In a controlled clinical trial with sample sizes of 65 patients in each group, statistically significant differences of 20 points can be detected on all eight SF 36 scales.

Conclusions—All eight scales of the SF 36 questionnaire show high reliability when used to monitor health in groups of patients, and at least four scales possess adequate reliability for use in managing individual patients. Further studies are required to test the feasibility of implementing the SF 36 and other outcome measures in routine clinical practice within the health service.

Introduction

In an era of escalating healthcare costs and scarce resources, purchasers and providers of health care require information that will allow them to “estimate as best they can the relation between medical interventions and health outcomes.” Only then will they be able to achieve their stated goals of efficient and high quality medical care. The kind of information systems required will need to incorporate measures of outcome that are valid; reliable; responsive to clinically significant changes in health over time; and, above all, quick and easy to administer in a routine clinical setting. Few outcome measures currently available for routine use satisfy all these criteria.

The short form 36 (SF 36) health survey questionnaire is a shortened version of a battery of 149 health status questions used in the RAND Corporation study of health insurance in the United States and was developed as a potential tool for monitoring patient outcomes in a busy clinical setting. The questions were selected to produce a questionnaire that could be completed in under 10 minutes while retaining the validity and reliability of the longer parent questionnaire. The SF 36 questionnaire measures three aspects of health: functional status, wellbeing, and overall evaluation of health using eight separate scales (table 1). The responses to the questions on each scale are summed to provide eight scores between 0 and 100. Because it is a general measure, the questionnaire may be used to compare health status both among patients with the same condition and between patients with different conditions. It may also be administered to general populations to see how a particular

Table 1 SF 36 health survey questionnaire scales

<table>
<thead>
<tr>
<th></th>
<th>No of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Functional status</td>
<td></td>
</tr>
<tr>
<td>(a) Physical functioning</td>
<td>10</td>
</tr>
<tr>
<td>(b) Social functioning</td>
<td></td>
</tr>
<tr>
<td>(c) Role limitations attributable to physical problems</td>
<td>4</td>
</tr>
<tr>
<td>(d) Role limitations attributable to emotional problems</td>
<td>3</td>
</tr>
<tr>
<td>II</td>
<td></td>
</tr>
<tr>
<td>Wellbeing</td>
<td></td>
</tr>
<tr>
<td>(a) Mental health</td>
<td>5</td>
</tr>
<tr>
<td>(b) Energy and fatigue</td>
<td>4</td>
</tr>
<tr>
<td>(c) Pain</td>
<td>2</td>
</tr>
<tr>
<td>III</td>
<td></td>
</tr>
<tr>
<td>Overall evaluation of health</td>
<td></td>
</tr>
<tr>
<td>(a) General health perception</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>35*</td>
</tr>
</tbody>
</table>

*The 36th question, which asks respondents to compare their health now with that one year ago, is not included within these eight scales.
condition causes health to depart from a “healthy standard.” Providing that the criteria of validity, reliability, and responsiveness are fully satisfied, the SF 36 questionnaire, as part of a routine outcomes monitoring system, has enormous potential for bringing about improvements in health care. For example, in the context of resource allocation, it may help purchasers to make difficult rationing decisions such that priority is given to those medical conditions for which treatment is of proven benefit (as in the Oregon experiment in the United States); it may facilitate audit of clinical practice and evaluation of new medical interventions in clinical trials; and it may inform the process of clinical decision making in managing individual patients by providing immediate feedback on patient functioning and wellbeing during the doctor-patient consultation.

Strong evidence for the clinical and psychometric validity of the SF 36 questionnaire in patient populations was provided by two recent studies in the United States and a study in the United Kingdom. In the United Kingdom study, for example, the SF 36 questionnaire was able to show appreciable differences in self reported health between the general population and patients with four common conditions. This included highly significant differences for patients with varicose veins, a condition for which treatment has been accorded low priority by some purchasers of health care in the United Kingdom.

Two approaches can be taken to test reliability: internal consistency, where similar questions within a scale are assessed for the extent to which they are correlated with each other and with the overall total score, and test-retest, where two measurements separated by some interval of time are assessed for differences. Controversy exists over the relative merits of both procedures. “Test-retest reliability has been put forward as the more rigorous approach when the instrument under investigation is to be used as an evaluative measure to detect changes in health status over time.” A major problem with test-retest, however, is its tendency to underestimate reliability if true change has occurred — that is, if patients actually get worse or improve between measurements. Although the internal consistency of the eight SF 36 scales has been well documented in both patient and general populations (C A McHorney et al, unpublished data), only one study, which used a general population sample, reported the test-retest reliability. If the SF 36 questionnaire is to be used to assess changes in health status over time its test-retest reliability should be demonstrated in patient populations.

We report the results of two studies in which both internal consistency and test-retest reliability were assessed, the first in more than 1700 patients presenting with four common conditions (low back pain, suspected peptic ulcer, menorrhagia, and varicose veins) and the second in a study of just over 570 patients attending a gastroenterology outpatient clinic.

**Methods**

**Sampling and Data Collection**

**Study 1**

Between March and June 1991 we identified patients in Grampian region presenting with one of four common conditions: low back pain, menorrhagia, suspected peptic ulcer, and varicose veins. The patients were identified in one of two ways: from all referral letters to outpatient departments in Grampian and by general practitioners from four large training practices in Grampian; these patients being included only if the general practitioner did not refer them to a specialist during the recruitment period of the study. A questionnaire including an anglicised version of the SF 36 questionnaire and sociodemographic questions was sent to the patients in general practice within two weeks of their initial consultation and to the referred patients before their first outpatient appointment. Patients not wishing to take part in the study were asked to return their questionnaire uncompleted. Reminders were sent to non-responders after two weeks and again after four weeks. In order to assess test-retest reliability a sample of patients returning a questionnaire was sent a retest questionnaire after two weeks. This included an additional question which asked, “Since you last completed a questionnaire, would you say that your health has improved, got worse, or stayed the same?”

**Study 2**

All patients attending the gastroenterology outpatient clinic in Aberdeen Royal Hospitals Trust during April 1993 were asked to complete a similar questionnaire, containing the SF 36 instrument, in the clinic waiting area before seeing the doctor. In order to assess test-retest reliability, a subsample of patients with ulcerative colitis was sent a retest questionnaire one week later, including three additional questions asking whether their physical, mental, or general health had improved, got worse, or stayed the same since they had last completed a questionnaire.

**Statistical Analysis**

We used two methods to assess reliability: internal consistency and test-retest reliability. Internal consistency was assessed using Cronbach’s α, which measures the correlations between questions that make up a scale at a single point in time. Alpha coefficients were calculated for each of the eight SF 36 scales based on responses to the initial questionnaire. An α coefficient exceeding 0.5 is considered acceptable when comparing groups of patients, although coefficients exceeding 0.9 have been recommended when making comparisons between individual patients or assessing change in scores in an individual patient over time. In order to demonstrate test-retest reliability it is necessary to show consistency between questionnaires for those patients reporting no change in health. The difference between the scores of these patients for both questionnaires was calculated, and the
standard deviations of those differences were used to calculate 95% confidence intervals for the differences in SF36 score for individual patients.\(^2\)

To make a thorough comparison of the two methods of assessing reliability the \(\alpha\) coefficients could be used to construct 95% confidence intervals of the differences. The measurement of any patient’s true score is a random variable with a variance which is equal to the product of the variance of the scores and one minus the reliability coefficient.\(^3\) Therefore when two independent measurements are made for all patients who reported no changes in their health, at two different points in time, 95% confidence intervals of their individual measurement differences could be constructed using statistical first principle analysis.\(^3\) The results of both methods of reliability assessment and the limits of the corresponding 95% confidence intervals for an individual’s SF36 score differences are presented separately for the two study populations.

Finally, to illustrate how reliability affects the precision of the SF 36 questionnaire – that is, its statistical power to detect clinically important differences in health status between different groups of patients, sample size estimates were calculated.\(^2\) Estimates of sample sizes required to detect small to large group differences in mean SF 36 scores, as defined by Ware and colleagues,\(^2\) were calculated for two kinds of study design: a randomised controlled clinical trial with comparisons between repeated SF 36 assessments over time and a before and after comparison of SF 36 scores in a single patient group.

**Results**

**RESPONSE RATE**

In study 1 a total of 1787 patients were identified, 800 by their general practitioners and 987 from referral letters, of whom 236 failed to respond and 193 refused to take part; 41 questionnaires were returned by the post office. Thus 1317 of 1746 correctly identified patients took part in the study, giving a final response rate of 75.4%. Their ages ranged from 16 to 86 years (mean 42.7 years), and 870 (66.1%) were female. Compared with those returning a questionnaire, non-responders were significantly younger (mean age 39.9 years; \(p < 0.01\)) but did not differ significantly in terms of sex, clinical condition, or source. A total of 775 retest questionnaires were posted to patients. Of the 710 (91.6%) returned completed, 414 (58.3%) reported no change in health status since returning the previous questionnaire.

In study 2, 573 patients were identified in the clinic, of whom, 549 (95.8%) patients returned a questionnaire. Their ages ranged from 12 to 86 years (mean 48.1 years), and 344 (62.7%) were female. Retest questionnaires were posted to 70 patients with ulcerative colitis. Of the 63 (90.9%) returned completed, 48 (53%) reported no change in physical, mental, or general health since completing the last questionnaire.

**STATISTICAL ANALYSIS**

In study 1 (the general practice and referred patients) estimates of reliability ranged from 0-80 (social functioning) to 0-92 (physical functioning), and from 0-66 (role limitations attributable to emotional problems) to 0-93 (physical functioning) with the internal consistency and test-retest methods respectively (table 2). The limits of the 95% confidence intervals for an individual patient’s SF 36 score differences ranged from ±20.3 (physical functioning) to ±45.2 (role limitations – emotional problems) based on the Cronbach’s \(\alpha\) and from ±19 (physical functioning and general health perceptions) to ±65.7 (role limitations – emotional problems) using the standard deviation of score differences from the test-rest analysis. The highest discrepancy between the two methods was in their reliability estimates for role limitations attributable to physical problems and role limitations attributable to emotional problems.

In study 2 (the gastroenterology clinic patients), correlation coefficients ranged from 0-80 (social functioning) to 0.93 (physical functioning), and from 0.79 (role limitations attributable to physical problems) to 0.94 (physical functioning) using the internal consistency and test-retest methods respectively (table 3). The limits of the 95% confidence intervals for an individual patient’s SF 36 score differences ranged from ±20.7 (physical functioning) to ±41.1 (role limitations – emotional problems) based on the Cronbach’s \(\alpha\), and from ±12.0 (mental health)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Reliability estimates with internal consistency method (n = 1286 minimum)</th>
<th>Reliability estimates with test-retest method (n = 414 minimum)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\alpha) coefficient</td>
<td>Mean SF 36 score</td>
</tr>
<tr>
<td>Physical functioning</td>
<td>0.92</td>
<td>70.7</td>
</tr>
<tr>
<td>Social functioning</td>
<td>0.80</td>
<td>70.7</td>
</tr>
<tr>
<td>Role limitations: Physical</td>
<td>0.89</td>
<td>43.6</td>
</tr>
<tr>
<td>Emotional</td>
<td>0.86</td>
<td>57.0</td>
</tr>
<tr>
<td>Mental health</td>
<td>0.86</td>
<td>65.0</td>
</tr>
<tr>
<td>Pain</td>
<td>0.86</td>
<td>50.3</td>
</tr>
<tr>
<td>Energy and fatigue</td>
<td>0.86</td>
<td>47.0</td>
</tr>
<tr>
<td>General health</td>
<td>0.83</td>
<td>62.9</td>
</tr>
</tbody>
</table>

*Table 2: Internal consistency and test-retest reliability estimates for eight SF 36 scales in patients with low back pain, menorrhagia, suspected uterine, and ear, nose, and throat disorders.*
to ±46·1 (role limitations – physical problems) using the standard deviation of score differences from the test-retest analysis. The highest discrepancy between the two methods was in their reliability estimates for role limitations attributable to physical problems and for mental health.

Finally, the effect of different degrees of reliability on study sample sizes required to detect differences in mean SF 36 scores for two types of study design was calculated. The sample size estimates required to detect small to large differences for a randomised controlled clinical trial with comparisons between repeated SF 36 assessments over time are shown in table 4. To detect a large difference of 20 points on all eight scales a sample size of 64 is required. The sample sizes required to detect the same differences for a before and after study design, with a comparison of SF 36 scores in a single patient group are shown in table 5. To detect a large difference of 20 points on all eight scales a sample size of 22 is required.

### Discussion

The internal consistency of the eight scales making up the SF 36 questionnaire has been reported in at least six published papers. However, this form of reliability testing relies on a single administration of the questionnaire, which may lead to "an optimistic interpretation of the true reliability of the test." A formal assessment of SF 36 questionnaire test-retest reliability was reported in only one study, in which respondents were selected at random from two general practitioner lists and thus represented the general population. In this paper we report on a comprehensive assessment of SF 36 reliability in two patient based studies.

In both studies the estimates of reliability based on internal consistency were remarkably similar, and all were within the range of estimates reported elsewhere. Appreciable differences between the test-retest and internal consistency methods of assessment were seen in study 1, but only for the two role limitation scales (in favour of the internal consistency method), and in study 2 for the mental health scale (in favour of the test-retest method). For example, the limits of the 95% confidence intervals around an individual patient’s SF 36 score difference on the mental health scale in study 2 were twice as wide when generated using internal consistency compared with the test-retest limits. It is worth noting here that in study 2 the time interval between administration of the questionnaire was only one week, and patients were included in the analysis only if they had reported no change in physical, mental, or general health. In study 1 the interval between administrations was two weeks, and patients were asked only whether their health had changed.

Despite the above mentioned differences between the two methods of reliability assessment, both methods produced similar results for most of the SF 36 scales. However, the accuracy of internal consistency estimates are generally likely to be higher than that of the test-retest method since, as in both of our reported studies, estimates of internal consistency are usually based on much larger samples. Test-retest requires the administration of an additional questionnaire, and is therefore usually undertaken on a smaller sub-sample of the main survey.
In summary, the widely held view that internal consistency, as a form of reliability assessment, may lead to an optimistic interpretation of the true reliability of the test has not been consistently supported by the results of our studies. These findings are in accordance with research on reliability in tests of mental ability; Lord and Novick, for example, found that test-retest methods may overestimate or underestimate reliability, depending on the time interval between administrations, and concluded that ‘‘internal consistency provides an accurate approximation to the reliability if the test under investigation is either reasonably homogeneous, or of substantial length and not speeded.’’

Tables 4 and 5 were constructed from the most conservative reliability estimates from study 1. The findings in these tables indicate that all eight SF 36 scales show very high levels of reliability when used to monitor health status and health outcome in groups of patients, so that even with sample sizes of around 65 patients in each group in a controlled clinical trial for example, statistically significant differences of 20 points can be detected on all eight scales. With sample sizes of only 30 patients in each group, statistically significant differences of 20 points are detectable on six of the eight scales. These data may be used as a guide by those wishing to estimate sample size for their studies, but other design factors may also need to be taken into account. If the SF 36 questionnaire is used to assess health status and health outcome as part of individual patient management however, our findings suggest that some of the SF 36 scales are not reliable enough to be of practical value. For instance, if a patient obtains a score of 50 on the scale measuring role limitations attributable to emotional problems, from the results obtained in our two studies, that patient’s true score could be anywhere between 9–91, and 15–85, based on test-retest results from studies 1 and 2 respectively. Individual scores obtained on the scale measuring role limitations attributable to physical problems also have wide confidence intervals. However, the reliability of scales measuring bodily pain and social functioning is probably adequate for detecting clinically important differences between individual patients or between sequential measurements on the same patient.

For example, if a patient with an arthritic hip scored 10 on the SF 36 scale measuring bodily pain, after hip replacement that patient would need a score of over 36 or 31 (based on results from studies 1 and 2 respectively) in order to register a significant improvement at the 5% level. The remaining scales – physical functioning, mental health, energy/fatigue, and general health – are sufficiently reliable to be able to detect even smaller differences. The same arthritic patient scoring 10 on the SF 36 physical functioning scale before surgery would need to achieve a postoperative score of over 26 or 25 (based on studies 1 and 2 respectively) to register a statistically significant improvement at the 5% level.

In conclusion, for a measure of health status and health outcome to be suitable for routine use in the NHS in a wide variety of clinical settings, it must provide information that is valid, reliable, responsive to change, quick to administer, and easy to collect. Our findings suggest that the SF 36 questionnaire is reliable enough to be used for monitoring health in groups of patients, and at least four of its scales possess adequate reliability for use in managing individual patients. In the accompanying article (p 186) we have also been able to show that the SF 36 questionnaire is responsive to clinical changes in health status over time. When this evidence is considered together with other published data on validity and reliability we believe it is sufficient to conclude that the SF 36 questionnaire result satisfies the requirements of a routine outcome measure. Further studies are now required to test the feasibility of implementing this and other outcome measures in routine clinical practice within the health service.

We thank the staff of Inverurie, Portlethen, Rubislaw Place, and Westhill practices, and the staff of the gastroenterology unit and medicine units of the Aberdeen Royal Infirmary and Aberdeen City General Hospital, in particular John Masson, Jeremy Grimshaw, Dean Phillips, and Johann Courts for the data collection and Elizabeth Russell for helpful comments. This research and the Health Services Research Unit are both funded by the Chief Scientist Officer of the Scottish Office Home and Health Department; however, the opinions expressed are those of the authors, not the NOHIS.

11 Garratt AM, Ruta DA, Abdalla MI, Buckingham JK, Russell IT. The SF 36 health survey questionnaire: an outcome measure suitable for routine use within the NHS. BMJ 1993;306:1440–4.
26 Garratt AM, Ruta DA, Abdalla MI, Russell IT. The SF 36 health survey questionnaire. II. Responsiveness to changes in health status for patients with four common conditions. Quality in Health Care 1994;3:186-92.
Reliability in two patient based studies.

D A Ruta, M I Abdalla, A M Garratt, A Coutts and I T Russell

*Qual Health Care* 1994 3: 180-185
doi: 10.1136/qshc.3.4.180

Updated information and services can be found at:
http://qualitysafety.bmj.com/content/3/4/180

**Email alerting service**

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

**Notes**

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/