Journal of Epidemiology and Community Health 1995;49:102-105

The short form 36 health status questionnaire: clues from the Oxford region's normative data about its usefulness in measuring health gain in population surveys

Sue Ziebland

Abstract

Objectives **– To determine the potential of the short form 36 health status questionnaire (SF-36) for indicating changes in the health status of a general population by examining the recently published normative data.**

Design **– The sensitivity of the SF-36 was tested through hypothesising two dramatic changes in health status whereby (i) the scores of people in social class V are improved to the level of social class I, and (ii) the scores of men and women aged 55 to 64 are altered to the level of current 45 to 54 year olds. The size of the effect measured by the SF-36 was considered.**

Results **– Small to moderate effects were evident when SF-36 mean scores for social class V were increased to the level of social class I, and primarily negligible effects were apparent on all domains but physical function for the postulated "10 years of age" improvement.**

Conclusion **– The SF-36 may be a useful measure for detecting changes in health status in homogenous treatment groups, but the variation in responses in a general population make it an inadequate tool for assessing the diffuse impact of health interventions directed at the whole community.**

(*J Epidemiol Community Health* 1995;**49**:102-105)

Health status questionnaires which measure the patients' views of their health have been increasingly used in treatment trials and recommended for routine monitoring of health.[1 2] The considerable attention which has been paid to reliability, validity, and sensitivity,[3 4] has meant that these may now be employed as an informative adjunct to traditional clinical measures in an ever growing catalogue of conditions.[5 6] The admission of the patient's account of their health, albeit in a highly structured and quantifiable form, into the arena of clinical practice has been an important and generally welcome development.

In recent years health status questionnaires have also been included as part of health and lifestyle and needs assessment surveys of the general population in the UK. Thus, in addition to descriptive data about those behaviours which are understood to be health related, structured accounts of self reported health in terms of pain, physical mobility, energy, emotional health, activities of daily living, social integration, and so on may be gained. If the pattern of responses suggests that a representative sample has been achieved, then surveys provide the potential to analyse scores on self reported health domains by such variables as gender, age, area of domicile, and ethnic group. One of the benefits of using a standard instrument is the availability of comparative data. Population norms for the Nottingham health profile have long been available[7] and data for the medical outcomes study SF-36[8] have recently been provided by the Oxford regional health and lifestyle survey.[9 10]

In this paper a third potential use of health status questionnaires will be examined: as an indicator of the changing health of a general population. In an evolution which parallels that in the health service, evaluators of urban development initiatives have begun to seek more direct outcome measures than such physical indicators as house building, road construction, and industrial development. The ultimate outcome is widely acknowledged to be quality of life, and manifestations of this elusive concept are sought through health status as well as educational attainment, resident perceptions, crime levels, and migration (Marion Headicar, personal communication). District health authorities with the responsibility for purchasing "health gain" will need to monitor the impact of services which are delivered through a wide range of channels including the acute sector, health promotion, community care, screening, and immunisation programmes. Some interventions are clearly more open to evaluation than others. Routinely collected indicators are available, yet these may be criticised for their insensitivity, as are mortality data, or considered unreliable, as are statistics on health service utilisation, which are notoriously subject to variations in the supply of services. They are also rather too narrow to "measure the potential benefits of health care interventions that can influence a wide number of variables such as physical mobility, emotional wellbeing, social life, and overall wellbeing."[9]

One way of attempting to track the impact of urban development, or a district's purchase of health services, may be through repeated population surveys designed to observe changes in perceived health status. Although many studies have included health status questionnaires

Camden and Islington Health Authority, 110 Hampstead Road, London NW1 2LJ
S Ziebland

Correspondence to:
GPRG,
University of Oxford,
Department of Public Health and Primary Care,
Gibson Building,
Radcliffe Infirmary,
Oxford OX2 10HE
Ms S Ziebland.

Accepted for publication May 1994

as outcomes measures in treatment trials, and a number of health and lifestyle surveys have included them as a snapshot indicator of the population's health, little is known about the practicality of using them as a measure of the impact of interventions on the whole community. Nevertheless, longitudinal population studies are already underway with designs which include the short form 36 health status questionnaire (SF-36) as an outcome measure.[11] If the effects of interventions on the health of the community are to be assessed, then the measures which are used must be shown to be sensitive to changes within the population. This is very different from detecting positive and negative effects in relatively homogenous groups undergoing similar treatments, which is the basis on which analysis of sensitivity has usually been conducted.

Studies including health status questionnaires as outcome measures for population based interventions are at relatively early stages, hence conclusive evidence of their sensitivity is not yet available. The potential appeal of a measure which is believed to assess the impact of such interventions should probably not be underestimated. This paper reports on a secondary analysis of the recently published normative data for the SF-36. The data will be examined for clues to the probable responsiveness of the questionnaire when included as an outcome measure in repeated population assessments.

Methods

CHARACTERISTICS OF THE SF-36

The SF-36 is a generic, self completion health status questionnaire with eight distinct dimensions (physical functioning, social functioning, role limitations (physical and emotional), general mental health, energy, bodily pain, and general health perceptions). A score is calculated for each of the dimensions by summing responses to individual items and converting, by a scoring algorithm, to a scale from 0 (poor health) to 100 (good health). Originally developed in America, in recent months a number of British studies have been published, adding weight to claims that the SF-36 is acceptable to respondents, reliable, and valid.[10 12]

Any questionnaire has internal characteristics such as the wording of the items, the scales for responses, and method of scoring, which will influence its sensitivity to change.[12 13] Items which invite the respondent to rate their health in comparison with those around them, such as the general health perceptions item on the SF-36: "I am as healthy as anyone I know", are destined to be inadequate reflectors of any general improvements or deteriorations in a population. This may explain why the differences in general health perception scores tend to be smaller across age group or social class than they are on other domains which are less overtly comparative in their focus. An additional problem, pointed out by Hunt and McKenna, is that questionnaires have a limited shelf life since "both the items and the language may have less relevance with time".[14] The "slipperiness" of language may not only hamper longitudinal studies, but can confound surveys which aim for broad targets without due care and attention.

SELECTION OF CROSS SECTIONAL COMPARISONS

In the absence of any completed longitudinal data sets the population norms which have recently been published from the Oxford health and lifestyle survey may be considered, although the nature of the analysis is inevitably constrained by the characteristics of the published data. If clues to the ability of the SF-36 to reflect health gain in the general population are to be obtained from cross sectional data, the choice of comparison groups should reflect the social patterning of health as characterised by sizeable differences in health status scores. Published data include a breakdown of SF-36 domain scores by sex, age group, manual/non-manual occupation, and social class (coded using the Registrar General's 1991 classification).[10] From these, the largest differences in mean scores were between the extremes of occupational class, which accords with well-documented observations on the social ordering of morbidity and mortality.[15] A very similar pattern is evident when the scores of those in the youngest (18–24) and oldest (55–64) categories are compared. Between adjacent age groups more modest differences are apparent.

In this paper two separate, and undeniably dramatic, transformations in the populations' health are hypothesised and the ability of the SF-36 to detect the magnitude of the impact assessed. These are (i) the mean reported health status of people in social class five being improved to that of social class one and (ii) the mean reported health status of 55 to 65 year olds being improved to that of men and women ten years their junior.

MEASURING CHANGES

One of the difficulties in using health status questionnaires for repeat measurements is in interpreting the meaning of raw change scores.[16 17] With a large sample even very minor differences will achieve significance. Overstatement of statistical significance will occur when an intervention programme is applied on an aggregate (such as a school, workplace, or community) and the analysis is based on individual observations which are, nevertheless, characterised by clusters of subgroups.[18]

Lydick and Epstein[17] emphasise the importance of clarifying the perspective of the research before interpreting changes in quality of life measures. "If we are speaking about the impact of an intervention on a population, perhaps we should not talk of clinical significance, but of public health significance or economic significance. A change in a measure which has been calibrated to be meaningful in population terms would not be expected to have the same relevance on an individual basis." Similarly, health status measures which have been de-

signed to assess the impact of specific medical interventions may not be suitable for use in the general population.

STATISTICAL ANALYSIS

Effect sizes have been recommended for comparing and translating changes in scores on health status measures.[19] Kazis *et al* have shown the use of effect sizes in identifying changes which are statistically useful in preference to the rather less discriminating criteria of statistical significance. Simple to calculate, an effect size presents the difference between a baseline and follow up mean score, adjusting for the spread of responses around the sample's mean score by dividing by the standard deviation of the responses before the intervention. In a group of respondents with broadly similar characteristics the standard deviations may be expected to be smaller than in a group combining people who are variously affected by physical, mental, and environmental limitations.

An effect size of 1·00 is equivalent to a change of one standard deviation in the sample. As a benchmark for assessing the relative magnitude of a change, Cohen[20] identified an effect size of 0·20 as small, one of 0·40 as moderate, and 0·80 as large.

Table 1 Effect sizes for hypothetical change in mean health status (as measured by domains of the SF-36) for men and women of social class V to that of social class I

	Social class V Mean (SD) (n=236) (min)	*Social class 1 Mean (SD) (n=384) (min)*	*Difference*	*Effect size*
Physical function	84·3 (21·3)	93·4 (11·7)	9·1	0·42
Social function	85·7 (21·3)	91·0 (16·7)	5·3	0·25
Role – physical	82·8 (33·0)	89·9 (25·2)	7·1	0·22
Role – emotional	79·7 (34·5)	87·3 (26·5)	7·4	0·21
Mental health	70·8 (20·0)	76·6 (14·7)	5·8	0·29
Energy	58·7 (20·3)	63·7 (18·8)	5·0	0·25
Pain	78·6 (23·2)	88·2 (16·2)	9·6	0·41
General health	70·3 (21·2)	75·1 (17·8)	4·8	0·23

Table 2 Effect sizes for hypothetical change in mean health status (as measured by domains of the SF-36) for men of 55 to 65 years of age, to that of men aged 45 to 54 years

	Men 45–54 y Mean (SD) (n=815) (min)	*Men 55–65 y Mean (SD) (n=619) (min)*	*Difference*	*Effect size*
Physical function	87·9 (17·4)	80·0 (22·1)	7·9	0·36
Social function	89·8 (18·7)	86·9 (22·6)	2·9	0·13
Role – physical	87·6 (28·3)	78·8 (36·1)	8·8	0·24
Role – emotional	85·7 (29·5)	85·8 (29·9)	−0·1	0·00
Mental health	76·0 (16·7)	78·0 (17·3)	−2·0	−0·12
Energy	62·9 (19·9)	62·9 (20·3)	0·0	0·00
Pain	81·8 (22·2)	78·8 (23·6)	3·0	0·13
General health	72·0 (20·1)	68·1 (22·9)	3·9	0·17

Table 3 Effect sizes for hypothetical change in mean health status (as measured by domains of the SF-36) for women of 55 to 65 years of age, to that of women aged 45 to 54

	Women 45–54 y Mean (SD) (n=918) (min)	*Women 55–65 y Mean (SD) (n=618) (min)*	*Difference*	*Effect size*
Physical function	84·8 (18·3)	74·8 (23·5)	10	0·43
Social function	87·0 (20·8)	85·9 (22·6)	1·1	0·05
Role – Physical	82·4 (32·0)	76·6 (36·9)	5·8	0·16
Role – emotional	80·8 (33·6)	83·3 (32·5)	−2·5	−0·01
Mental health	73·2 (18·2)	74·4 (18·5)	−1·2	−0·06
Energy	59·4 (20·3)	59·0 (21·4)	0·4	0·02
Pain	77·4 (22·3)	75·0 (25·1)	2·4	0·10
General health	73·1 (19·9)	68·0 (22·0)	5·1	0·23

Results

Between 1991 and 1992 the Oxford regional health and lifestyle survey collected lifestyle, demographic, and health status data. A self completion booklet, which included the SF-36 health status questionnaire, was mailed to a random sample of 18 to 64 year olds drawn from the computerised registers of the family health service authorities within the Oxford health region. A response rate of 72% provided 9332 respondents. A method exists for the substitution of values for missing data, but this was not used in this survey because of the large sample size and the objective of providing normative data. The following secondary analysis is based on the published data[9 10] for the Oxford survey.

Tables 1, 2, and 3 present hypothetical tests of the sensitivity of the normative SF-36 data from the Oxford survey. The data used in these tables are the mean scores and standard deviations of scores on domains of the SF-36, broken down by social class, sex and age group. Table 1 expresses the effect sizes calculated for domains of the SF-36 for a hypothetical improvement in the self reported health of people in social class V to the current level of social class I. There can be no doubt that such a change would be seen as a spectacular achievement for health services. However, the calculation of effect sizes would identify the domains of physical function and pain as only moderately changed, while all other domains would register as small effects.

Tables 2 and 3 show that if the mean scores of men and women aged between 55 and 65 were improved to that of people 10 years their junior, the effect sizes identified by changes in SF-36 scores would be mostly negligible. Small effects for physical role for men and for women's general health perceptions would be apparent and only the difference in physical function scores would be considered moderate.

Discussion

The use of health status questionnaires in treatment trials is well established and, notwithstanding the limitations of busy clinical

practice, are being increasingly used as a welcome addition to routine patient monitoring. Generic questionnaires, among which the SF-36 is currently attaining the highest profile in the UK,[9 10 14 21] may augment the descriptive faculties of health and lifestyle surveys. To what extent the availability of scores on health status questionnaires actually enhances the assessment of the *health needs* of a population is debatable. Donovan *et al* argue that scores on the Nottingham health profile provide too obscure a clue to health care requirements to be of much use in supporting purchaser decisions.[22]

In the search for a compact solution to the problem of assessing the impact of policies on a community, longitudinal assessment of self reported health status may seem appealing as an outcome measure. Studies are in progress which include health status measures in this capacity. Effect sizes have been identified as the most widely used of the distribution based interpretations of changes in health related quality of life measures.[17] However, the implication of the effect size calculations shown here is that the SF-36, at least, is unlikely to be able to give any useful indication of the type of changes which can be expected to occur within the health of a population as a result of community wide interventions. This analysis suggests that, even if policies did improve the health status of those in social class V to that of social class I, the SF-36 would not be equal to the task of measuring the outcome.

At this point it must be stressed that there is no evidence that the evaluation of changes in the general population was ever the intention of the designers of the SF-36. It is not really surprising that by amalgamating the undeniably "chalk and cheese" nature of self report health status in the population, one would be unable to detect clearly the effects of interventions. However, the responsibility for evaluating population based interventions often does not lie with those who are most aware of the original purposes and inherent limitations of a questionnaire. Surveys certainly have their methodological place, but if repeat assessments are made using insufficiently sensitive measures, there is a distinct danger that either the results will be inconclusive or, more alarmingly, that wholly inappropriate conclusions will be drawn about the effectiveness of policies. This is not to say that interventions should not be evaluated, but simply that suitable methods should be used. Small scale, carefully focussed, research employing imaginative designs will often provide far more interpretable data about the impact of a policy than that discernible through the broad sweep of the survey.

Use of instruments such as the SF-36 as outcome measures may be entirely appropriate when applied to a relatively homogenous group of patients. Although evidence of responsiveness has yet to be demonstrated within the UK,[10] there are encouraging indications and it is certainly not the object of this paper to dampen enthusiasm for such use.

The examples of hypothetical changes chosen here were deliberately extreme. Much discussion might surround the interpretation of more minor changes, but any intervention which succeeded in transforming the health of the lowest social class to that of the highest would be recognised as quite remarkably effective. Some of the effects of ageing may be less clearcut, but physical (if not mental) functioning is reported as deteriorating between the decades of middle age.

The explanation for the poor responsiveness shown in this paper seems to lie in the variability of responses on SF-36 domains within the general population. Hence, the standard deviations of the mean scores limit the responsiveness, even in what would be acknowledged to be a sensational change. This would not augur well for measurement of the more modest goals which are the apotheosis of current policies.

I am grateful to Lucie Wright and Crispin Jenkinson of the Health Services Research Unit at the University of Oxford and to Ian Basnett and Brenda Chipperfield of Camden and Islington Health Authority for their encouraging comments. Thanks also to the anonymous referee who provided helpful comments on an earlier draft.

1 Spilker B, Molinek F, Johnson K, *et al.* Quality of life bibliography and indexes. *Med Care* 1990;**28**(Suppl):DS1–DS77.
2 Wolfe F, Pincus T. Standard self report questionnaires in routine clinical and research practice – an opportunity for patients and rheumatologists. *J Rheum* 1991;**18**:643–6.
3 Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. Importance of sensitivity to change as a criterion for selecting health status measures. *Quality in Health Care* 1992;**1**:89–93.
4 Guyatt G, Deyo R, Charlson M, *et al.* Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiology* 1989;**42**:403–8.
5 Bowling A. *Measuring health: A review of quality of life measurement scales.* Milton Keynes: Open University Press, 1992.
6 Wilkin JE, Hallam L, Doggett M. *Measures of need and Outcome for Primary Care.* Oxford: OUP, 1992.
7 Hunt SM, McEwan J, McKenna SP, *et al.* Measuring health status: A new tool for clinicians and epidemiologists. *J R Coll Gen Pract* 1985;**35**:185–8.
8 Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF36): conceptual framework and item selection. *Med Care* 1992;**30**:473–83.
9 Jenkinson C, Coulter A, Wright L. Short form 36 (SF36) health survey questionnaire: normative data for adults of working age. *BMJ* 1993;**306**:1427–40.
10 Jenkinson C, Wright L, Coulter A. *Quality of life measurement in health care. A review of measures, and population norms for the UK SF-36.* Oxford: Health Services Research Unit, 1993.
11 Nuffield Institute for Health, Health Outcomes Clearing House. Personal communication, June 1993.
12 Ziebland S, Fitzpatrick R, Jenkinson C. Tacit models of disability underlying health status instruments. *Soc Sci Med* 1993;**37**:69–75.
13 Ziebland S, Fitzpatrick R, Jenkinson C, Mowat A, Mowat A. Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the health assessment questionnaire (HAQ) and modified HAQ. *Ann Rheum Dis* 1992;**51**:1202–5.
14 Hunt SM, McKenna SP. Measuring patients' views of their health. (Letter). *BMJ* 1993;**307**:125.
15 Davey Smith G, Bartley M, Blane D. The Black report on socioeconomic inequalities in health 10 years on. *BMJ* 1990;**301**:373–7.
16 Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measurement in health care: 1 Applications and issues in assessment. *BMJ* 1992;**305**:1074–7.
17 Lydick E, Epstein RS. Interpretations of quality of life changes. *Quality of Life Research* 1993;**2**:221–6.
18 McKinley J. The promotion of health through planned sociopolitical change: challenges for research and policy. *Soc Sci Med* 1993;**36**:109–17.
19 Kazis L, Anderson JA, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;**27**:S178–S189.
20 Cohen J. *Statistical power analysis for the behavioural science.* New York: Academic Press, 1977.
21 Garratt AM, Ruta D, Abdalla M, *et al.* The SF36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? *BMJ* 1993;**306**:1440–4.
22 Donovan J, Frankel S, Eyles J. Assessing the need for health status measures. *J Epidemiol Community Health* 1993;**47**:158–62.