Making sense of ambiguity: evaluation of internal reliability and face validity of the SF 36 questionnaire in women presenting with menorrhagia

Crispin Jenkinson, Viv Peto, Angela Coulter

Abstract

Objective—To determine the face validity and internal reliability of the short form 36 (SF 36) health survey questionnaire in women presenting with menorrhagia.

Design—Postal survey of women recruited by their general practitioners followed by interviews of a selected subsample.

Patients—348 women who had consulted their general practitioner with excessive menstrual bleeding and completed questionnaires after treatment. 49 women selected from this group were interviewed in depth about their health status, and requested to complete the SF 36 questionnaire.

Main measures—Subjective accounts of functioning and wellbeing as measured by the eight scales of the SF 36 questionnaire.

Results—Data from the postal survey indicated that the "general health perceptions" and "mental health" scales of the SF 36 questionnaire had lower internal reliability coefficients than documented elsewhere. In the follow up interviews several questions on the SF 36 questionnaire were commented on as inappropriate or difficult to answer for patients with heavy menstrual bleeding.

Conclusions—Some questions on the SF 36 questionnaire were difficult to answer for this group of patients. Such problems can adversely effect the validity of the measure. It is suggested that comments of patients upon measures such as the SF 36 questionnaire could both determine the appropriateness of such measures for given studies and influence questionnaire design.

(*Quality in Health Care* 1996;5:9-12)

Keywords: short form 36 (SF 36) health survey questionnaire, internal reliability, face validity, menorrhagia.

Introduction

In recent years considerable interest has centred on the use of health status questionnaires in evaluating medical treatment. These questionnaires, completed by patients, can provide invaluable data for assessing treatment outcomes and are a potentially vital addition to traditional outcome measures of morbidity and mortality.¹⁻⁴ Advocates of this approach argue that such data can provide very accurate

assessments of the health of individual subjects or populations and the benefits and harm that may result from medical care.⁵

The ideal outcome of treatment is a return to the normal quality of life for a given age and medical condition.^{6,7} Subjective health measures can be used as one part of the evaluation of the outcome of treatments, supplementing traditional endpoint measurements, such as morbidity and mortality. For the information to be useful the data gained from subjective health measures must conform to certain standards. Thus much of the research undertaken on measures of subjective wellbeing documents their psychometric and statistical validity and reliability; detailed consideration of what those who complete these questionnaires actually understand by the questions is rarely considered.⁸

The issue of how questions are interpreted or reinterpreted is known in psychometrics as "face validity." Streiner and Norman, in their well regarded text on issues of questionnaire design and validation, claim that:

"Face validity simply indicates whether, on the face of it, the instrument appears to be assessing the desired qualities. The criterion represents a subjective judgement based on a review of the measure itself by one or more experts, and rarely are any empirical approaches used."⁹

Thus they are claiming that the designers of questionnaires, who are "experts," are the arbiters of what makes sense on a questionnaire. However, perhaps it should not be experts (psychometricians or clinicians) who assess the face validity of measures, but a representative group of people in the population to be targeted with a given questionnaire. If a questionnaire is to be completed by clinicians then a representative sample of such clinicians should be asked to complete it before its general adoption. If the questionnaire is to be completed by patients then clearly it is they who should be the arbiters of its face validity. This may seem unremarkable, but it often seems to have been ignored in the design and validation of questionnaires on health status.

One measure that has been advocated for routine use in the National Health Service is the short form 36 health survey questionnaire (SF 36 questionnaire). This measure is brief and has excellent validity and reliability. In this paper we record the use of the SF 36 questionnaire in patients presenting with menorrhagia;

Health Services
Research Unit,
University of Oxford,
Department of
Public Health and
Primary Care,
Radcliffe Infirmary,
Oxford OX2 6HE
Crispin Jenkinson,
deputy director
Viv Peto, research officer

King's Fund Centre for
Health Services
Development,
126 Albert Street,
London NW1 7NF
Angela Coulter, director

Correspondence to:
Dr Jenkinson

Accepted for publication
7 September 1995

assess the internal reliability of the questionnaire for these patients; and present data on the difficulty some patients had in answering it.

Subjects and methods

RECRUITMENT

Patients presenting with heavy menstrual bleeding were recruited from 73 general practices in the Oxford Regional Health Authority. The patients were told about the study and given an information sheet and questionnaire, to be completed after the visit. Those eligible for entry were aged 30–49 years, presenting with either a first or subsequent complaint of heavy menstrual bleeding regardless of the severity of symptoms. All women who returned completed questionnaires were entered into the study. Eighteen months later a follow up questionnaire, including the SF 36 questionnaire was posted to respondents, who were asked to complete and return it in a prepaid envelope. Two reminder letters were sent if necessary, together with further copies of the questionnaire. Inclusion in the study in no way altered the patients' medical treatment.

This paper is solely concerned with the internal reliability and face validity of the SF 36 questionnaire. The data reported are based on the returns from the follow up questionnaire. A subsample of these respondents participated in a semistructured interview about their health care and any issues raised by the use of the SF 36 questionnaire.

SF 36 QUESTIONNAIRE

The SF 36 questionnaire consists of 36 questions within eight multi-item scales (table 1). There is a further unscored single item about health change over the past year. For each scale item scores are coded, summed, and transformed on to a scale from 0 (worst possible health state measured by the questionnaire) to 100 (best possible health state). The questionnaire was designed in the United States and hence minor modifications to the wording of six scales items have been made to make it acceptable in the United Kingdom.¹⁰

ANALYSIS

Data from the postal survey

Internal reliability refers to the ability of dimensions of a questionnaire to tap the phenomena they claim to be considering. The internal reliability of the eight scales of the SF 36 questionnaire was assessed with Cronbach's α statistic.¹¹ It has been suggested that an α value of 0.5 or above is acceptable, although Nunnally recommends values of 0.7 and

above.¹² Internal reliability estimates calculated for patients presenting with menorrhagia were compared here with estimates calculated for a sample of the general population from the Oxford region with the cut off points recommended by Nunnally.¹² Full details of the postal study of patients presenting with menorrhagia are published elsewhere.^{13–15}

Patients' perception of face validity

A subsample of respondents to the final postal questionnaire were asked if they would be willing to undertake an interview about their experiences of health care. The first 50 who agreed, were interviewed. They were asked to complete an interviewer administered copy of the SF 36 questionnaire with the specific request that they orally highlight any difficulties they had in understanding or completing the questions. Their comments were taped, transcribed, and subjected to content analysis, which entailed reading through the data and identifying recurrent themes or those areas which were emphasised as important by the interviewees.

Results

Four hundred and eighty three patients were recruited into the study and the 425(88%) who returned the first questionnaires were followed up after treatment at 18 months. In all, 348(82%) patients returned follow up questionnaires. Table 2 shows internal reliability of the different scales of the SF 36 questionnaire for this group of patients. These results were compared with those from the Oxford healthy lifestyle survey (OHLS) in which a randomly selected sample of the general population completed the SF 36 questionnaire. Results from the patients with menorrhagia showed a much lower internal reliability than those from the Oxford Survey, especially for the general health perceptions and mental health scales.

INTERVIEW SAMPLE

Fifty women were interviewed; 49 (of whom 35 still reported heavy periods) completed the SF 36 questionnaire in the presence of the interviewer. These women made comments on how they understood the questions. On average, these respondents took seven (range 2–13) minutes to complete the SF 36 questionnaire.

WHAT PATIENTS SAID ABOUT THE SF 36 QUESTIONNAIRE

Many of the patients with menorrhagia who were interviewed were unsure whether they

Table 1 SF 36 questionnaire scales

	No of items
General health perceptions	5
Physical functioning	10
Social functioning	2
Role limitations (physical)	4
Role limitations (emotional)	3
Mental health	5
Energy of vitality	4
Pain	2
SF-36 change item: Change in health over the last year	1

Table 2 Internal reliability of scales of the SF 36 questionnaire for a random sample of the general population and for a sample of patients presenting with menorrhagia (sample sizes vary due to incomplete questionnaires)

Scale	Population sample α coefficient	Patients with menorrhagia α coefficient
Physical functioning	0.90(8883)	0.89(340)
Social functioning	0.76(9219)	0.68(346)
Role limitations (physical)	0.88(9151)	0.83(347)
Role limitations (emotional)	0.80(9159)	0.80(346)
Mental health	0.83(9014)	0.50(347)
Energy or vitality	0.85(9089)	0.90(347)
Pain	0.82(9214)	0.85(348)
General health perceptions	0.80(9085)	0.51(343)

were intended to complete the SF 36 questionnaire with reference to their health in general or with specific reference to their periods. Despite the fact that the questionnaire specifically asks respondents to complete it with reference to general health, 12 of the patients expressed confusion as to what exactly the questionnaire meant by "health in general." Respondents found it difficult to disentangle their overall view of their health from that of their problems due to menorrhagia or that of the side effects of operations to reduce it. One woman remarked that she knew that she was in a study to evaluate the impact of treatment on her menorrhagia. Therefore she assumed all questions related to this aspect of her health. For example:

"I know it does say general (health), but, funnily enough, in my mind, I was still thinking with regard to the operation, because the study is on that."

Similarly, another woman, who had a heavy cold, did not wish to report on her general health because it had nothing to do with her menorrhagia. On reading the first item of the SF 36 questionnaire, which asks patients to assess their health in general, she remarked:

"Do you think your health is excellent? At the moment I don't feel it is at all. It's difficult to see. I haven't been well because of this sort of virus. I had it all over Christmas, and it's come back to me after about a week or so. So, during the past four weeks I didn't go out...because I didn't feel well, but that's nothing to do with this (menorrhagia)."

Questions with a time frame added to the difficulties in completing the questionnaire in relation to menorrhagia or health in general. For example, all items on the mental health scale refer to problems in the past month whereas those in the general health scale refer to an unspecified time period. Both time frames are ambiguous for women presenting with menorrhagia. Ten respondents commented on this problem. As one respondent stated, she had a problem with questions asking about the past month because she thought that her health was fine but her periods made her feel terrible. She reported having severe premenstrual tension associated with depression and instances of bad temper for two weeks before heavy bleeding. However, despite these demands she found the questionnaire difficult to complete as it requested general accounts of her health. Except for the problems of her periods she definitely perceived her health as generally good. She was aware that this seemed contradictory but could think of no consistent manner in which to complete many of the items. Thus, respondents found themselves drawn between answering questions relating to their general health, which mostly they seemed to regard as good, and answering in relation to their menorrhagia, which caused problems only at certain times.

Discussion

Subjective health measures such as the SF 36 questionnaire have been advocated by many as

essential to monitoring medical care because they incorporate the views of patients. Until recently, scientists and physicians have assessed the impact of interventions on traditional end points; quality of life and subjective health status have been inferred indirectly.¹⁶ Data based on patients as derived from health status measures, were rarely collected. However, with the new emphasis on "patient based outcomes assessment," the use of such questionnaires has been advocated as potentially appropriate for everything from population based health assessment to clinical trials, as well as for the routine evaluation of medical care, both for groups⁴ and individual patients.¹⁷ Indeed an RT-2000 scanning system, which scans a specially printed version of the SF 36 questionnaire and provides print-outs of individual patient scores, is now available.¹⁸

Such data have the potential to transform medical care, in that they could provide important insights into the impact of interventions on the wellbeing of those receiving them. Furthermore, they could influence clinical decisions. It is imperative therefore, that such data accurately reflect what they claim to measure. The α coefficients reported here, for example, suggest that SF 36 questionnaire data gained from patients with menorrhagia would not be sufficiently reliable for individual clinical decision making, and would be potentially misleading in evaluating change over time in individual women presenting with this condition. Certainly the reliability coefficients on the mental health and general health perceptions scales in this group of patients are far below those recommended as acceptable for use at the level of individual patients.^{12 17 19}

Table 3 shows the content of these two scales. It is apparent that both the nature of the questions and the time frame could be a problem for patients with an illness with symptoms not present continuously. For example, the items of the general health perceptions scale ask patients to assess their health in general. For someone in perfect health such a question is not a problem. Similarly, for someone with a longstanding illness the answer may seem perfectly clear. However, menorrhagia is

Table 3 Items in the general health perceptions and mental health scales of the SF 36 questionnaire

Health perceptions scale:

In general, would you say your health is:
Excellent/Very Good/Good/Fair/Poor?

Please choose the answer that best describes how true or false each of the following statements is for you:

I seem to get ill more often than other people
I am as healthy as anybody I know
I expect my health to get worse
My health is excellent

Respondents select from "Definitely true/Mostly true/Not sure/Mostly false/Definitely false" for each of the above items

Mental health scale:

How much time during the past month:

Have you been a very nervous person?
Have you felt so down in the dumps nothing could cheer you up?
Have you felt calm and peaceful?
Have you felt downhearted and low?
Have you been a happy person?

Respondents select from "All of the time/Most of the time/A good bit of the time/Some of the time/A little of the time/None of the time" for each of the above items

characterised by symptoms that are not continuously present. Furthermore, although the condition may be distressing, most patients are aware that it is neither life threatening nor inevitable that the symptoms will persist for their entire lives. Thus, questions about general health contain a certain ambiguity: on the one hand such patients visit their doctors because they wish to have certain symptoms alleviated, but on the other hand regard their health as excellent. This point was highlighted by respondents who answered the SF 36 questionnaire at interview.

These results suggest that due caution must be exercised when selecting questionnaires as they may pose questions that respondents, owing to the distinctive nature of their symptoms, find both ambiguous and misleading. Simply because a measure of subjective health status performs well in general population surveys or with certain groups of patient, this cannot be taken to imply its appropriateness for all groups or conditions. To make data from measures such as the SF 36 questionnaire interpretable it is essential that pilot trials of the questionnaires are carried out on all groups on which they are to be used. Such studies will provide crucial information on the appropriateness of a given measure for any given population.

- 1 Wasson J, Keller A, Rubenstein L, Hays R, Nelson E, Johnson D, and the Dartmouth Primary Care COOP Project. Benefits and obstacles of health status assessment in ambulatory settings: the clinician's point of view. *Med Care* 1992;30(suppl):MS42-9.
- 2 Jenkinson C. Evaluating the efficacy of medical treatment: possibilities and limitations. *Soc Sci Med* 1995;41:1395-401.
- 3 Geigle R, Jones SB. Outcomes measurement: a report from the front. *Inquiry* 1990;27:7-13.
- 4 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.
- 5 Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measures in health care. I. Applications and issues in assessment. *BMJ* 1992;305:1074-7.
- 6 Ware J. Measuring patients' views: the optimum outcome measure. SF-36: a valid, reliable assessment of health from the patient's point of view. *BMJ* 1993;306:1429-30.
- 7 Silver GA. Paul Anthony Lembock, MD, MPH: a pioneer in medical care evaluation. *Am J Psychiatry* 1990;80:342-8.
- 8 Donovan JL, Frankel S, Eyles JD. Assessing the need for health status measures. *J Epidemiol Community Health* 1993;47:158-62.
- 9 Streiner DL, Norman GR. *Health measurement scales: a guide to their development and use*. Oxford: Oxford University Press, 1989.
- 10 Brazier JE, Harper R, Jones NMB, O'Cathain A, Thomas KJ, Usherwood T, et al. Validating the SF-36 health survey questionnaire: new outcome measures for primary care. *BMJ* 1992;305:160-4.
- 11 Cronbach L. Coefficient alpha and the structure of tests. *Psychometrika* 1951;16:297-334.
- 12 Nunnally JC. *Psychometric theory*. 2nd ed. New York: McGraw, 1978.
- 13 Coulter A, Peto V, Jenkinson C. Quality of life and patient satisfaction following treatment for menorrhagia. *Fam Pract* 1994;11:394-401.
- 14 Coulter A, Peto V, Doll H. Patients' preferences and general practitioners' decisions in the treatment of menstrual disorders. *Fam Pract* 1994;94:67-74.
- 15 Jenkinson C, Peto V, Coulter A. Measuring change over time: a comparison of results from a global single item of health status and the multi-dimensional SF-36 health status survey questionnaire in patients presenting with menorrhagia. *Quality of Life Research* 1994;3:317-21.
- 16 Leighton-Reed J. The new era of quality of life assessment. In: Walker S, Rosser R, eds. *Quality of life assessment: key issues in the 1990's*. Lancaster: Kluwer, 1993:00-00.
- 17 Neil N. The interstudy quality edge 2: the presentation of health status data. Excelsior, MN: Interstudy, 1992.
- 18 Ware J, Snow K, Kosinski M, Gandek B. *SF-36 health survey manual and interpretation guide*. Boston: Health Institute (New England Medical Center), 1993.
- 19 Helmstadter GC. *Principles of psychological measurement*. New York: Appleton-Century-Crofts, 1964.