The Design and Assessment of Questionnaires in Clinical Research

S M Saw, T P Ng

ABSTRACT

Questionnaires are one of the most commonly used tools for data collection in clinical research. Despite its simplicity and convenience of use, the design of questionnaire instruments that accurately measure health status and their determinants is nevertheless a difficult and challenging task. We review the two most important issues which are reliability and validity. Reliability can be defined as the degree to which a measure gives 'consistent' or 'reproducible' values when applied in different situations. Validity refers to the extent in which the true value of a variable is correctly measured by the instrument. For different types of questionnaire measurement instruments, specific issues of content, construct and criterion validity should be appropriately addressed. Accuracy in questionnaire-based measurement in clinical studies is achieved by paying attention to the relevant specific issues of reliability and validity during development and testing of such questionnaires.

Keywords: Reliability, Validity, Instruments, Questionnaires

Singapore Med J 2001 Vol 42(3):131-135

INTRODUCTION

Questionnaires are one of the most commonly used techniques for collecting health-related information in clinical studies because of their ease and simplicity of use. Although not all modalities of information can be collected with questionnaire techniques, a wide and unique range of health-related information is nevertheless possible. In clinical research and evaluation of clinical practice, the information of interest includes health outcomes such as illness severity, adverse events related to care, functional status, and satisfaction with care.

Although the use of questionnaire information to measure health status and risk factors are commonplace because of its convenience, the most important and challenging aspect of the questionnaire methodology lies in the design and development of questionnaires that accurately measures health status and risk factors of interest[1]. In the present paper, we review the important aspects in the design and development of reliable and valid questionnaire tools in clinical research.

DEVELOPMENT AND TESTING OF QUESTIONNAIRES

The questions that are to be included in a questionnaire are determined by consideration of the health outcomes and risk factors that are likely to explain variations in health status. This may require comprehensive literature searches to ascertain the conceptual and operational definitions used in different studies. The relevant factors that are likely to play possible roles as confounders and effect modifiers may also be identified from these literature reviews. A review of questionnaires used in similar past studies would be useful when designing the new questionnaire. It should be noted, however, that the known validity and reliability of such published questionnaires should be ascertained as much as possible, as it should not be assumed that the same level of validity and reliability will apply in different population groups. A rating scale may have validity in one context, e.g. hospital based care, yet may not be valid in another context, e.g. community-based care.

Questionnaires should ideally be pilot tested with a small convenience sample of people and refined such that the questionnaire is simple to answer and yet gives accurate data. Attention to layout, coding, order and type of questions (open or close ended) are important factors in its final design[2,3]. The final form of the questionnaire will also depend on whether it is administered by an in person interview, a telephone interview or self-administered by the participant.

RELIABILITY AND VALIDITY

The most important consideration in the design and administration of a questionnaire is that it must be able to measure *accurately* what it is designed to measure. The *'accuracy'* of the data obtained from a questionnaire has two components: reliability and validity. *Reliability* is defined as the degree to which a measure gives

Department of Community, Occupational, and Family Medicine
National University of Singapore
16 Medical Drive (MD 3)
Faculty of Medicine
16 Medical Drive
Singapore 117597

S M Saw, MBBS, MPH, PhD
Assistant Professor

T P Ng, MD, MFPHM
Associate Professor

Correspondence to:
Dr Saw Seang Mei
Tel: 874 4976
Fax: 779 1489

Table I. Summary of studies of the reproducibility of questionnaires.

Author (year)	Study population	Type of questionnaire	Administration	Time interval between two questionnaires	Results
Willett et al (1985)	173 female registered nurses aged 34 -59 years in Boston	Semi-quantitative food frequency questionnaire	Self-administered	One year	Intraclass correlation coefficient of 0.63 for total calories
Tsubono et al (1995)	492 residents of a rural Japanese town aged 40 to 69 years	Food frequency	Self-administered	Several intervals ranging from two weeks to five years	Median Spearman rank correlation coefficients at 2 weeks and 5 years were 0.62 and 0.28 respectively
Aaron et al (1995)	100 adolescents aged 15 to 18 years in Pittsburgh	Physical activity	Self-administered	Two intervals of one month and one year	Spearman rank correlation coefficients at one month and one year were 0.79 and 0.66 respectively
Westerdahl et al (1996)	670 women in Sweden	Assessment of melanoma risk	Self-administered	One interval ranging from one to three years	Kappa statistics ranging from 0.4 for naevi on the right arm to 0.95 for smoking

'consistent' or 'reproducible' values when applied in different situations, such as on different occasions on the same individual (test-retest reliability), or on the same individual by different interviewers (inter observer reliability), or when a number of similar question-items are intended the same entity (Inter-item consistency). 'Reliability', 'consistency' and 'reproducibility' are often used to mean the same thing when they are all different. Ideally, one would like a reproducible questionnaire instrument to give values that vary little under such circumstances. This reduces measurement variation ('background noise') and contributes to greater 'precision' in statistical estimates of the measure. *Pari passu*, an instrument which gives consistently the same values for a measurement variable will tend to give 'statistically significant results' than one which gives inconsistent values.

Validity refers to the extent to which the true value of a variable is correctly measured by the instrument. Whereas reliability may be compared to the ability of a marksman to get all his shots closely bunched up, validity is the marksman's ability to get all his shots closely at the bull's eye. The validity of an instrument is affected by the reliability. If there is poor reliability, validity will be reduced; if the shots are widely scattered, they will not be close to the bull's eye.

RELIABILITY AND ITS MEASUREMENTS

The reliability of the questionnaire may be assessed by administering the questionnaire at two different points of time and seeing the degree of variation that occurs ('test-retest reliability'). *Intra-subject variation* may occur where the measurement variable within the subject varies with the time of the day. For example, the dietary fat intake varies with the time of the day. Also, *intra-observer and inter-observer variation* may occur if the same questionnaire is administered by face-to-face as compared with a telephone interview.

The reproducibility of food frequency questionnaires and physical activity questionnaires has been reported in many studies. Table I is a summary of studies of the reproducibility of questionnaires[4-7]. The reproducibility of the questionnaire varies with the type of information collected and the time interval of administration of the two questionnaires[4]. The studies conducted in the United States, Japan, and Sweden showed good reproducibility of questionnaires that were administered at intervals ranging from two weeks to a few years.

With questionnaires which measure conditions or states represented by categorical variables (e.g. 'disease/non-disease', 'mild/moderate/severe'), reproducibility is most commonly and appropriately assessed using the *Cohen's kappa statistic*[8]. The *kappa* measures the agreement above and beyond the amount of agreement which would be expected by chance alone. A *kappa* of 0 to 0.2 indicates slight agreement, 0.21 to 0.4 indicates fair agreement, 0.41 to 0.60 moderate agreement, 0.61 to 0.80 substantial agreement and 0.81 to 1.00 perfect agreement.

For continuous measures, the *Pearson product-moment correlation coefficient* is often used, or the corresponding *Spearman rank correlation coefficient* for skewed data distribution. The *Pearson correlation coefficient* measures the strength of co-variability and not

exact value agreement between two measurements[9,10]. *Pearson correlation coefficient* may be misleadingly high, even though there is a systematic bias between the two measurements. The *intraclass correlation coefficient* is often the preferred statistical index for the exact agreement between two measurement variables[11].

VALIDITY AND ITS MEASUREMENTS

Validity is assessed by comparing the observed value against the 'true value'. This is ideally done by comparing it against a 'gold standard' measure which supposedly gives values closer to the 'truth', if such a external criterion measure is available. This is called '*criterion validity*'. For example, the validity of 'hard' entities such as 'diabetes', or 'tobacco smoke exposure' is easily assessed using criterion measures such as the oral glucose tolerance test, or urinary cotinine levels respectively.

Content validity

'*Content validity*' is concerned with how well the question-items correspond to the concept (or 'domain', 'construct') of what is being measured. For example, the Beck Depression Inventory (BDI)[12] is based on the traditional concept of depression, which includes 'biological symptoms' such as loss of appetite and sleep disturbance, which could be attributed to other illnesses, whereas the Hospital Anxiety and Depression (HAD) scale[13] is based on a revised conceptual map of depression, without its associated physical symptoms.

Content validity is assessed using qualitative techniques. In the development of multi-item rating scales, the content validity of the questionnaire items may be examined by using an expert panel, focus groups or in-depth interviews with respondents. Focus groups may be formed with a range of subjects representing typical extremes (for example very dissatisfied and very satisfied patients) and discussions should be guided by open-ended questions designed to elicit common and typical responses based on real experiences or perceptions by the subjects.

Construct validity

The '*Construct validity*' of a questionnaire applies when a single content or a single criterion cannot be determined. Typical examples are 'intelligence', 'personality', 'quality of life', or 'patient satisfaction'. Construct validity is present when a measurement scale is related to other measures predicted by theory or empirical observations. An example of construct validity is the well known construct of 'neuroticism' by Eysenck which has been validated by a body of observed inter-relations between variables which were predicted by the construct.

Unlike criterion validity, evidence for construct validity cannot be obtained from a single study, but from a number of inter-related studies. Construct validity has two components: *convergent validity* demonstrates association with measures that are or should be related, and *divergent validity* demonstrates a lack of association with measures that should not be related. There are different measures of 'abnormal illness behaviour'. One would therefore expect that these different measures would be correlated (convergent validity'), and they would show lack of correlation with measures of other independent constructs such as 'neuroticism' (divergent validity'). Similarly, one would expect a high correlation between 'physical functioning' dimension in the SF36 quality of life scale with the 'physical mobility' dimension in the Nottingham health profile scale. There should be much less correlation between 'physical functioning' in the SF36 scale with 'social isolation' in the Nottingham scale[14].

Assessment of validity - sensitivity and specificity

The *sensitivity, specificity, positive predictive, and negative predictive value* are the commonly used indices of criterion validity of the information from the questionnaire. *Sensitivity* is defined as the probability that the individual with the trait is correctly identified by the questionnaire as having the trait. For example, a questionnaire on smoking habits is sensitive if individuals who smoke (as identified by carbon monoxide-haemoglobin as the 'gold standard' instrument) are correctly identified by the questionnaire as smokers. *Specificity* is the probability that the individuals without the trait are correctly identified by the questionnaire that the trait is not present. Thus, for the same questionnaire on smoking habits, the specificity is high if the individuals who do not smoked as identified by carbon-monoxide-haemoglobin are non-smokers. The *positive predictive value* is the probability that a positive test will correctly identify people with the specified trait. An example of a questionnaire with a high positive predictive value is a questionnaire on alcohol consumption where individuals who are identified as regular drinkers are in fact, regular drinkers in real life (as identified by the 'gold standard' test). *Negative predictive value*, on the other hand, is the probability that a negative test will accurately identify people without the trait. The positive and negative predictive values depend on the prevalence of the measurement trait in the population. Further information on the *clinical utility* of the questionnaire measurement are provided by indices such as the *likelihood ratio*[15]. The likelihood ratio is simply defined as the ratio true positive rate (sensitivity) versus the false positive rate (1-specificity). The optimum cut-off value

Table II. Summary of studies of the validation of questionnaires in epidemiologic research.

Author (Year)	Study population	Type of questionnaire	'Gold standard' instrument used for validation	Results
Willett et al (1985)	173 Boston area female registered nurses aged 34 to 59 years	Food frequency	Four one-week diet records	Intraclass correlation coefficient of 0.37
Munger et al (1992)	44 women aged 55 to 69 years in Iowa	Food frequency	Five 24-hour dietary recalls	Median adjusted Pearson correlation coefficient = 0.45 for macronutrients
Rimm et al (1991)	127 male health professionals in Boston aged 40 to 70 years	Expanded food frequency	Two one-week diet records 6 months apart	Mean intraclass correlation coefficients for energy-adjusted nutrient intakes = 0.59
Elosua et al (1994)	187 Spanish men aged 20-60 years	Minnesota leisure time physical activity	Maximal treadmill exercise test	Spearman rank correlation coefficient of 0.57 between total activity metabolic index and exercise test duration
Aaron et al (1995)	100 adolescents 12 to 16 years in Pittsburgh	Physical activity	Four 7-day physical activity recalls	Spearman correlation coefficients of 0.55 to 0.83

of a questionnaire measurement variable to determine the most number of true positives and the least number of false positives is commonly determined from *receiver operating characteristics analyses.*

'Gold standards' in validation studies

Physical activity from a questionnaire may be compared to physical activity as measured by more objective measures such as basal heart rate or exercise fitness test[16]. Diet intake from a food frequency questionnaire may be validated against a series of one week diet records as in the Nurses' Health Study[4]. The VF-14 questionnaire was designed to measure functional impairment caused by cataract and it was validated against visual acuity and global self-rating of the overall amount of difficulty that patients had with their vision[17]. Both neurotic and depressive symptoms in self-report questionnaires were validated with a psychiatric interview and examination[18,19].

A summary of several validation studies in different populations[4,6,16,20,21] is described in Table II. The validation studies were conducted for food frequency and physical activity questionnaires in populations ranging in sample size from 44 to 187. The correlation coefficients were good ranging from 0.3 to 0.8.

Questionnaire bank

It is essential that questionnaire-based measurements be reliable and valid, as well as available for evaluation by other study researchers interested in the same area of clinical research. Questionnaires are often not rigorously tested and not subject to peer review for reliability and validity. If these questions have been vigorously pre-tested and validated in other studies, they should be made available for other researchers as appendices in published papers or stored in a questionnaire bank[22,23]. If details on the development and testing of the questionnaire are included in the paper as well, one would definitely have better quality of data. This will facilitate the development of new and better improved questionnaires to measure the various health outcomes.

CONCLUSION

Questionnaires are useful and common tools in clinical research. Questionnaires are not expensive and may be used to measure a large number of health outcomes including medical diseases of interest, functional status, and quality of life. Attention to the relevant specific issues of reliability and validity in the development and testing of the questionnaire should be done and they should be formally assessed and reported. The standardization of questionnaires will enable us to collect quality data which is essential for clinical research.

REFERENCES

1. Carmines EG, Zeller RA. Reliability and validity assessment. SAGE Publications; London, 1979.
2. Stone DH. How to do it: Design a questionnaire. BMJ 1993; 307:1264-6.
3. Helsing K, Comstock G. Response variation and location of questions within a questionnaire. Int J Epidemiol 1976; 5:125-30.
4. Willett W. Nutritional Epidemiology. Monographs in epidemiology and biostatistics. Oxford University Press, New York. 1990
5. Tsubono Y, Nisino Y, Fukao A, Hisamichi S, Tsugane S. Temporal change in the reproducibility of a self-administered food frequency questionnaire. Am J Epid 1995; 1231-5.
6. Aaron DJ, Kriska AM, Dearwater SR, Cauley JA, Metz KF, LaPorte RE. Reproducibility and validity of an epidemiologic questionnaire to assess past year physical activity in adolescents. Am J Epidemiol 1995; 142:191-201.
7. Westerdahl J, Anderson H, Olsson H, Ingvar C. Reproducibility of a self-administered questionnaire for assessment of melanoma risk. Int J Epid 1996; 25:245-51.

8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74.
9. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. The Statistician 1983; 32:307-17.
10. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 307-10.
11. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med 1990; 20,5:337-40.
12. Beck AT, Ward CH, Mendelson M, et al. An inventory for measuring depression. Arch Gen Psych 1961; 4:561-71.
13. Zigmond AS, Snaith RP. The Hospital and Anxiety Scale. Acta Psych Scand 1983; 67:361-70.
14. Garrat AM, Ruta DA, Abdalla MI, Buckingham JK, Russel IT. The SF36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? BMJ 1993; 306:1440-44.
15. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993; 39,4:561-77.
16. Elosua R, Marrugat J, Molina L, Pons S, Pujol E. Validation of the Minnesota leisure time physical activity questionnaire in Spanish men. Am J Epidemiol 1994; 139:1197-209.
17. Steinberg E, Tielsch J, Schein OD, Javitt J, Sharkey P, Cassard SD, et al. The VR-14. An index of functional impairment in patients with cataract. Arch Ophthalmol 1994; 112:630-8.
18. Peveler RC, Fairburn CG. Measurement of neurotic symptoms by self-report questionnaire: validity of the SCL-90 R. Psychol Med 1990; 20:873-9.
19. Riley WT, McCranie EW. The depressive experiences questionnaire: validity and psychological correlates in a clinical sample. J Pers Assess 1990; 54:523-33.
20. Munger RG, Folsom AR, Kushi LH, Kaye SA, Sellers TA. Dietary assessment of older Iowa women with a food frequency questionnaire: nutrient intake, reproducibility, and comparison with 24-hour dietary recall interviews. Am J Epidemiol 1992; 136:192-9.
21. Rimm E, Giovannucci E, Stampfer M, Colditz G, Litin L, Willett W. Reproducibility and validity of an expanded self-administered semiquantitative food frequency questionnaire among male health professionals. Am J Epidemiol 1992; 135:1114-9.
22. Gordis L. Assuring the quality of questionnaire data in epidemiologic research. Am J Epidemiol 1979; 109:21-4.
23. Massey JT, Moore TF, Parsons VL, Tadros W. Design and estimation for the National Health Interview Survey: 1985-1994. Vital Health Stat (2) 1989; 110. DHHS publication PHS:89-1384.