Methodology Series Module 9: Designing Questionnaires and Clinical Record Forms - Part II

Maninder Singh Setia

Abstract
This article is a continuation of the previous module on designing questionnaires and clinical record forms in which we have discussed some basic points about designing the questionnaire and clinical record forms. In this section, we will discuss the reliability and validity of questionnaires. The different types of validity are face validity, content validity, criterion validity, and construct validity. The different types of reliability are test-retest reliability, inter-rater reliability, and intra-rater reliability. Some of these parameters are assessed by subject area experts. However, statistical tests should be used for evaluation of other parameters. Once the questionnaire has been designed, the researcher should pilot test the questionnaire. The items in the questionnaire should be changed based on the feedback from the pilot study participants and the researcher’s experience. After the basic structure of the questionnaire has been finalized, the researcher should assess the validity and reliability of the questionnaire or the scale. If an existing standard questionnaire is translated in the local language, the researcher should assess the reliability and validity of the translated questionnaire, and these values should be presented in the manuscript. The decision to use a self- or interviewer-administered, paper- or computer-based questionnaire depends on the nature of the questions, literacy levels of the target population, and resources.

Key Words: Questionnaires, reliability, validity

Introduction
This article is a continuation of the previous module on designing questionnaires and clinical record forms. In the previous module, we discussed some basic points in designing the questionnaire and clinical record forms. In this section, we will discuss the reliability and validity of questionnaires.

Examples
Carlsson and coworkers (2017)
The authors assessed the inter- and intra-observer reliability of Hand Eczema Extent Score (HEES). Six dermatologists assessed the hand eczema in 18 patients twice (before and after lunch). The patients were hidden by a screen, and the dermatologists could only visualize the hands. Furthermore, the patients were also instructed not to wear any items that could help in identification on their hands (such as rings, bracelets, nail paints). Interestingly, the presence of tattoos was an exclusion criterion. They reported that the intra-observer reliability was good (intraclass correlation coefficient [ICC] 0.88–0.94). Their ICC for intra-observer reliability for the two observations by patients was 0.95. Thus, they concluded HEES is reliable to grade the extent of eczema.

Sarkar and colleagues (2016)
The authors developed a questionnaire for assessing quality of life in melasma patients in Hindi. They translated the original Melasma Quality Of Life Questionnaire (MELASQOL) in Hindi and evaluated the questionnaire properties. The steps in translations were (1) translation; (2) review by the expert committee; (3) back translation; (4) rereview by the committee; (5) pretesting; and (6) final revision. They reported the Cronbach’s alpha of Hindi-MELASQOL was 0.861; thus, they concluded that reliability was satisfactory. The authors also compared the correlation between Melasma Area Severity Index (MASI) and Hindi-MELASQOL. They reported that the Spearman’s correlation coefficient was 0.835. Their ICC for inter-observer reliability was 0.927. They concluded MELASQOL is reliable tool for assessing the quality of life.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as the author is credited and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

How to cite this article: Setia MS. Methodology series module 9: Designing questionnaires and clinical record forms – Part II. Indian J Dermatol 2017;62:258-61.

Received: April, 2017. Accepted: April, 2017.
correlation between these was 0.809 (P < 0.05); thus, they concluded that there was a significant positive correlation between MASI score and Hindi-MELASQOL.

The authors have also provided all the items in the manuscript. This questionnaire is useful for assessing QOL in Indian patients.

We urge the readers to read these manuscripts. They discuss practical methods for assessing the reliability and validity of questionnaires.

**Validity**

**What is validity?**

A test or questionnaire should measure what it is supposed to measure – this represents the validity of the test. Of all the definitions and explanations, I found the one provided by Aday to be complete and self-explanatory. Aday describes the validity of a questionnaire as “the degree to which there are systematic differences between the information obtained in response to the questions relative to: full meaning of the concept; related questions about the same concept; and theories or hypotheses about their relationship to the concepts” (Aday, 1996).

**Types of validity**

We will discuss different types of validity in this section. The researchers have to address these points while designing a new scale, instrument, or measurement questionnaire. It should be remembered that all the questionnaires and instruments are based on certain theoretical concepts and measure some specific outcomes.

For example, quality of life questionnaire in acne patients should measure various aspects of quality of life (as desired by the researcher). To achieve this, the researcher should ensure that individual items or questions in the questionnaire are framed in such a manner that they actually measure various aspects of quality of life. Furthermore, if the researcher wants to measure different aspects of quality of life (health, social, and economic), then the items should measure that very aspect for which they were designed. For this process, it is important that questionnaire is valid. The different types of validity have been presented below.

**Face validity**

This is a subjective assessment of the questionnaire or instrument. This assessment is usually done by a subject area expert. After designing the instrument, the researcher should evaluate if the instrument appears appropriate and has relevant items on the “face of it.”. This is called “face validity.”

**Content validity**

This is also another form of subjective assessment. In this type of evaluation, the researcher assesses whether the items in the questionnaire adequately measure the concepts they are supposed to.

For example, if the item is meant to assess the social problems faced due to acne? Is the question designed to assess social problems? Or does it measure, economic problems? The subject area expert will assess the item and provide the comments to the researcher on the content validity of the questionnaire.

Thus, as discussed face and content validity are often assessed by subject area and research experts.

**Criterion validity**

It should be remembered that the questionnaire has been designed to measure specific outcomes. Thus, the measurement from this questionnaire should match with the existing standard (often called the “gold standard”).

For example, if we have designed a short scale to assess facial hyperpigmentation, then the results from this scale should match with the results from a standard assessment tool (such as objective assessment using a Colorimeter or Mexameter).

This validity is called “criterion validity.” There are two types of criterion validity—“concurrent validity” and “predictive validity.”

The above example of facial hyperpigmentation scale may be considered as “concurrent validity.” Another example - we develop a scale to assess the time spent in sun and the study participants have to fill the questionnaire at the end of every week. As a part of the study, all the participants are also fitted with GPS locators (just an example!), and we measure the actual time from this GPS locator. The comparison of the scale and the GPS assessments will be an example of concurrent validity.

Predictive validity scale evaluates a condition in the present and predicts some event in the future. For example, we develop a scale to assess the clinical, demographic, and histological features of patients with psoriasis. They are given a score; based on this score, we would like to predict which patients will develop erythroderma. Of course, this event will be in the future. Thus, this form of validity is predictive validity of the scale.

**Construct validity**

In this form of validity, the researcher examines the items in the questionnaire with respect to some underlying hypotheses or “construct.”

Sometimes, concepts cannot be measured directly but need to be evaluated using multiple indicators or variables. For instance, pigmentation, hair width, may be assessed objectively using modern techniques. However, other assessments require elaborate questionnaires. Let us consider an example for this - we have to develop
a questionnaire to evaluate psychoneurocutaneous
conditions. For this questionnaire, we need to include
appropriate items/questions in the questionnaire.
Usually, these items are based on some underlying
hypotheses or constructs. Some of these items/questions
may evaluate physical health; some may evaluate mental
health, whereas others may assess other forms of stress.
It is important that responses to the items that are
supected to evaluate physical health are similar and
they should be different from responses to mental health
or stress. The former is called “convergent validity,” and
the latter is called “divergent validity.”

This form of validity is assessed using statistical methods
such as factor analysis or principal component analysis.
These are advanced techniques; hence, the researcher
should consult a statistician for these analyses.

Reliability
What is reliability?
If we design a questionnaire to measure a particular
concept, then it is important that this questionnaire
asseses the same concept over time on repeated testing.
Furthermore, the responses and outcomes should be
consistent and similar when the questionnaire is used by
different researchers. This “stability” of the questionnaire
is called reliability (Aday 1996).

Types of reliability
In the previous section, we have discussed different
forms of validity. In this section, we will highlight the
different types of reliability and how to assess it.

Test-retest reliability
As discussed earlier, the questionnaire or the instrument
should give similar responses when used over time,
unless there is an actual change in the outcome.
Let us consider the above questionnaire to assess
psychocutaneous diseases. We have three different types
of questions – those evaluating physical health, mental
health, and other forms of stress. If we administer the
questionnaire over two time points (with a difference
of 1 day), then the responses should be similar (unless
there has been a stressful event between these two time
points). The agreement between these two responses will
represent the test-retest reliability.

It is important that the time between two responses
should be appropriate. The difference should not be
so large that the change in the response is an actual
change in the response. For example, if we are measuring
the blood pressure, then the changes over a couple of
days may represent an actual change. If the agreement
is weak, we should not conclude that the measuring
instrument is not “reliable.” However, when we are
assessing physical health or mental health, there should
not be major changes in the response over the same
time period, unless there has been an acute physical or
a mental health event. Thus, in this case, the agreement
should be high for questionnaire to be “reliable.”

Intra-rater reliability
The responses to the items in the questionnaire should
be similar when the same interviewer administers it at
two or more time points. The agreement between these
two responses represents the inter-rater reliability.

Inter-rater reliability
The responses to the items in the questionnaire should
be similar when two different interviewers administer it.
The responses should not depend on the person who is
administering the questionnaire. The agreement between
these two responses represents the inter-rater reliability.

How do we assess reliability?
These different types of reliability are measured with
statistical tests such as Pearson’s correlation, Spearman’s
rank correlation, and ICC. Other tests include kappa
coefficient, Cronbach’s alpha, and Bland and Altman Plots.

A correlation value more than 0.70 is desirable
(where 1.00 denotes perfect correlation). However, in
some scenarios (particularly when the investigation
will influence management), a test-retest reliability of
more than 0.90 is required. Similarly, a kappa statistic
of 0.61–0.80 is considered substantial agreement,
and a value of 0.81–0.99 is considered almost perfect
agreement.

Additional Points
Once the questionnaire has been designed, the
researcher should pilot test the questionnaire. The
items in the questionnaire should be changed based
on the feedback from the pilot study participants and
the researcher’s experience. Some of the changes after
the pilot test can change the format or the language
of the questions; change the response patterns; remove
some questions; or add some new questions or items.
After the basic structure of the questionnaire has been
finalized, the researcher should assess the validity and
reliability of the questionnaire or the scale. It is possible
that a test with high reliability may have poor validity.
Thus, it is important to assess both these aspects for
each questionnaire.

The questionnaire can be self administered or
interviewer administered. A self-administered
questionnaire can be paper based or computer based.
The interviewer-administered questionnaire can be
face-to-face or over the telephone. In both these
scenarios, the researcher has the option to develop a
paper- or a computer-based questionnaire. The decision
to use a self- or interviewer-administered, paper-
or computer-based questionnaire depends on the nature of
the questions (individuals may be comfortable answering
personal questions in a self-administered questionnaire); literacy levels of the target population; and resources.

If an existing standard questionnaire is translated in the local language, the researcher should assess the reliability and validity of the translated questionnaire. These values should be presented in the manuscript.

Summary
In this section, we have discussed the reliability (test-retest, inter-rater, and intra-rater) and validity (face, content, criterion, and construct) of questionnaires. Once the questionnaire has been designed, the researcher should pilot test the questionnaire. The items in the questionnaire should be changed based on the feedback from the pilot study participants and the researcher’s experience. After the basic structure of the questionnaire has been finalized, the researcher should assess the validity and reliability of the questionnaire or the scale. The decision to use a self- or interviewer-administered, paper- or computer-based questionnaire depends on the nature of the questions, literacy levels of the target population, and resources.

Financial support and sponsorship
Nil.

Conflicts of interest
There are no conflicts of interest.

Bibliography
1. Aday, L.A., Designing & Conducting Health Surveys. Second ed. 1996, San Francisco, USA: Jossey-Bass.
2. Carlsson, A., Svensson, A., Anderson, C.D., Baranovskaya, I., Hindsen-Stenstrom, M., Holt, I., et al. (2017). Scoring of Hand Eczema: Good Reliability of Hand Eczema Extent Score (HEES). Acta Derm Venereol, 97(2), 193-197.
3. Cazzaniga, S., Lecchi, S., Bruze, M., Chosidow, O., Diepgen, T., Goncalo, M., et al. (2015). Development of a clinical score system for the diagnosis of photoallergic contact dermatitis using a consensus process: Item selection and reliability. J Eur Acad Dermatol Venereol, 29(7), 1376-1381.
4. Sarkar, R., Garg, S., Dominguez, A., Balkrishnan, R., Jain, R.K., & Pandya, A.G. (2016). Development and validation of a Hindi language health-related quality of life questionnaire for melasma in Indian patients. Indian J Dermatol Venereol Leprol, 82(1), 16-22.
5. Stephen B. Hulley, et al., Designing Clinical Research. Second ed. 2001, Philadelphia, USA: Lippincot Williams & Wilkins.
6. Streiner DL and Norman GR., Health Measurement Scales - a practical guide to their development and use. 2003, Oxford, UK: Oxford University Press.
7. Viera, A.J., & Garrett, J.M. (2005). Understanding interobserver agreement: The kappa statistic. Fam Med, 37(5), 360-363.