

## Validating Questionnaires

As we all know from the political polls that abound every fall, the answer you get depends very much on the question you ask. In order for a questionnaire to be useful, the data it produces must be trustworthy, i.e., we must know that the results are meaningful and can be applied more generally than to just the sample tested. Proving that trustworthiness for questionnaires involving subjective clinical endpoints is not trivial, and ensuring that the resulting data reflect the “truth” has spawned an entire field of study. This article will provide some insight into what that process entails, and why it is important.

The term “validation” has a variety of meanings in clinical research, the first and most obvious being the assessment of computer systems to ensure they function as expected. “Validation” is also the process by which any data collection instrument, including questionnaires, is assessed for its dependability. Validating questionnaires is somewhat challenging as they usually evaluate subjective measures, which means they can be influenced by a range of factors that are hard to control. In other words, a blood pressure machine can be assessed for accuracy and calibrated to ensure consistent readings. Obviously the SF-36 quality of life questionnaire cannot be similarly assessed.

That said, there are ways to evaluate the value of, or validate, a questionnaire. Validation involves establishing that the instrument produces data that are reliable and true. There are a number of ways to define this, some of which outlined below.

Reliability: the degree to which a questionnaire will produce the same result if administered again, or the “test-retest” concept. It is also a measure of the degree to which a questionnaire can reflect a true change.

Validity: the degree to which a questionnaire reflects reality. There are a number of different facets to validity.

Internal validity: the degree to which questions within an instrument agree with each other, i.e., that a subject will respond to similar questions in a similar way. It also affects the likelihood of producing false positives and false negatives.

External validity: the ability to make generalizations about a population beyond that of the sample tested.

Sensitivity: the degree to which the instrument can identify a true positive, e.g., accurately identify a person who does have the condition.

Specificity: similar to sensitivity, this is the degree to which the instrument can identify a true negative, e.g., correctly identify the people who do not have the disease. Sensitivity and specificity are another side of the coin from internal validity.

Statistical validity: this is related to internal validity, and assesses whether the differences in the questionnaire results between patient groups can appropriately be subjected to statistical tests of significance.

Longitudinal validity: whether a questionnaire returns the same results in a given population over time, assuming all else remains equal

Linguistic validity: whether the wording of the questionnaire is understood in the same way by everyone who completes it.

Discriminant validity: the ability of the questionnaire to detect true differences between groups, and detect no difference when there isn't one.

Construct validity: the ability of a measure to assess correctly a particular cause and effect relationship between the measure and some other factor.

"Validity" is not an absolute quality. It's a continuum, with a questionnaire being valid to a certain degree in certain circumstances, and researchers must decide (preferably before the validation study is run) what degree of validity is considered sufficient. The above categories also suggest that there are types of validity that relate to the internal validity of the questionnaire (are similar questions answered similarly), others that relate to the ability of the questionnaire to determine a given state in a patient (e.g., that it varies in alignment with the severity of the condition), and still others that involve the validity of comparing different groups on the basis of the questionnaire.

Each type of validity is distinct, meaning that a questionnaire can have one kind of validity but not another. Because of that, a questionnaire can never really be fully "validated." It can only be validated for *x* patient population, under *y* conditions, and so forth. This implies that it may not be appropriate, for example, to use a lymphoma quality of life questionnaire in a melanoma study if the questionnaire hasn't been validated for that particular population, unless it has been shown to be applicable to cancer patients generally.

The validity of the results can be impacted by more than just the design of the questionnaire itself. Some questionnaires must be administered by individuals who have been trained in survey administration generally, or that one in particular. Others can be administered by any experienced clinician, or nurse, or indeed completed by the patient. If an otherwise valid questionnaire is administered by the wrong individual, the results are compromised. Similarly, some instruments must be used in their original published form, and changing the layout to create a CRF may compromise the results. Results can also be compromised if the questionnaire is not completed at the expected times (either time or day or relative to some other event), or in the right setting.

The process of validating an instrument varies depending upon what aspect(s) of validity are being assessed. Generally it involves running a study that is designed to determine a specific kind of validity, although it is sometimes possible to add a validation arm onto a trial with other primary objectives. One way to check the validity of a questionnaire is to compare its results with results from more objective measures. For example, a questionnaire assessing a patient's perception of their chronic obstructive pulmonary disease (COPD) may be compared to measures of their lung function, and the results of each compared between groups of healthy subjects and ill patients. If the instrument has appropriate specificity, sensitivity and discriminant validity, one should see a good correlation between the lung functions of the more severely ill patients with "worse" scores on the questionnaire. The degree to which the differences in the scores vary in alignment with the lung function tests across the healthy and ill subjects is the measure of the validity of the instrument *at identifying patients who have COPD*.

If the same questionnaire was developed in the US in English, and researchers wanted to use it in Italy, it would need to be translated into Italian. The Italian version would then have to be tested to see whether it varied with degree of illness in the way the English one did, or at least in a reliable and predictable fashion. Of course, there may be cultural differences that may require changing the content of the instrument. "Walking the length of a city block" is generally understood in the US, but the concept is meaningless in rural France.

Establishing longitudinal validation is particularly relevant to clinical trials, in that determining the degree to which the use of an instrument repeatedly in a study affects the instrument results. On the one hand, in order to be able to draw conclusions from the results, the same instrument should be used throughout the study. For that to work it must be longitudinally valid. There is a well documented test-retest effect, however; the first time a subject completes a given questionnaire the results are independent. After that, the subject is no longer naïve to the questions, and their answers in the second questionnaire may be influenced by their memory of their prior experience. Part of the process of validating instruments used over time is statistically evaluating that relationship.

There are many cases in clinical research where no validated instrument exists for a given disease or population. That doesn't mean that the disease can't be assessed, or that questionnaires cannot be used. The protocol authors must decide how important that objective is to the study, and whether it is acceptable to have less trustworthy results for that objective. It suggests that unvalidated instruments should probably not be used for key efficacy or critical safety assessments until appropriate validation studies have been conducted. Alternatively, the company can consult with the regulatory authorities to ensure that their acceptance of the instrument prior to submission of the NDA.

As with so much in clinical research, there are no black and white rules where it comes to assessing the reliability and validity of questionnaires. Each drug development team must determine their study objectives and the best way of achieving those objectives. If that includes the use of questionnaires, they will need to assess what type and level of validation is sufficient for their purposes. As data managers, we can play a role in ensuring that those discussions happen before the study starts and it's too late to change.

## References

Cook TD, Campbell, DT. *Quasi-Experimentation; Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin Company, 1979  
Damato S, Bonatti C, Frigo V, Pappagallo S, et al. Validation of the Clinical COPD questionnaire in Italian language. *Health and Quality of Life Outcomes* 2005, 3:9