16

Surrogate Marker Validation in Mental Health

Tony Vangeneugden, Ariel Alonso Abad, Helena Geys, and Annouschka Laenen

16.1 Introduction

In this chapter, we describe how the framework for surrogate marker validation in clinical trials can easily be adapted and used to assess the so-called criterion validity of psychiatric symptom scales. This concept will be described further in this chapter (see also Laenen et al. 2004).

One feature of the psychiatric health sciences literature, devoted to measuring subjective states, is the daunting area of available scales (Steiner and Norman 1995). The development of scales to assess subjective attributes is not easy and subject to many controversial debates. One particular drawback, of course, lies in the fact that the filling-in of a scale may vary from one person to another. Because of the subjective nature of many of these scales, one may encounter scales that are not adequate to assess a particular concept. Therefore, whenever a mental health measurement scale is developed, translated, or used in a new population, its psychometric properties have to be assessed. Two important properties are reliability and validity.

Reliability consists in determining the extent to which the measurement is free from random error. This can be performed through analyzing internal consistency and reproducibility of the questionnaire. Internal consistency is the extent to which individual items are consistent with each other and reflect a single underlying construct. Essentially, internal consistency represents the average of the correlations among all the items in the instrument. Several measures that are often used to provide proof of internal consistency are: Cronbach’s alpha coefficient (Cronbach 1951), Kuder and Richardson (1953), and factor analysis. Intra-observer or test-retest reliability is the degree to which a measure yields stable scores at different points in time for patients who are assumed not to have changed clinical status on the domains being assessed. The calculation of intraclass correlation coefficients
(Fleiss and Cohen 1973, Deyo, Dierh, and Patrick 1991) is one of the most commonly used methods. For interviewer-administered questionnaires, the inter-observer reliability is the degree to which a measurement yields stable scores when administered by different interviewers, rating the same patients. The calculation of interclass correlation coefficients is also one of the most commonly used methods. In classical test theory, the outcome of a test is frequently modeled as

$$X = \tau + \varepsilon,$$

(16.1)

where $X$ represents an observation or measurement, $\tau$ is the true score, and $\varepsilon$ the corresponding measurement error. It is further assumed that the measurement errors are mutually uncorrelated as well as with the true scores, and under this assumption they obtain

$$\text{Var}(X) = \text{Var}(\tau) + \text{Var}(\varepsilon).$$

(16.2)

The reliability of a measuring instrument is defined as the ratio of the true score variance to the observed score variance, i.e.,

$$R = \frac{\text{Var}(\tau)}{\text{Var}(\tau) + \text{Var}(\varepsilon)}.$$  

(16.3)

For interviewer-administered questionnaires, the inter-observer reliability is the degree to which a measurement yields stable scores when administered by different interviewers, rating the same patients. Also here the interclass correlation coefficient is commonly used.

The validity of a questionnaire is defined as the degree to which the questionnaire measures what it purports to measure. This can be performed through the analysis of content, construct, and criterion validity. Content validity can be defined as the extent to which the instrument assesses all the relevant or important content or domains. Also the term face validity is used to indicate whether the instrument appears to be assessing the desired qualities at face. This form of validity consists of a judgment by experts in the field. Construct validity refers to a wide range of approaches which are used when what we are trying to measure is a “hypothetical construct” (e.g., anxiety, irritable bowel syndrome, ...) rather than something that can readily be observed. The most commonly used methods to explore construct validity are extreme groups (apply instrument for example to cases and non-cases), convergent and discriminant validity testing (correlate with other measures of this construct and not correlate with dissimilar or unrelated constructs), and multitrait-multimethod matrix (Campbell and Fisk 1959). Criterion validity can be divided into two types: concurrent validity and predictive validity. With concurrent validity we correlate