The validation of an instrument to diagnose depression: Beyond the yes/no question

A valid instrument is essential for any activity, be it in the clinical, educational or research field. Of relevance to the validation of an instrument in a new language is the extent of the benefits that it will provide. Herein lies the merit of the study: “Translation and validation of brief patient health questionnaire (BPHQ) against DSM IV as a Tool to diagnose major depressive disorder (MDD) in Indian patients,” published in this issue of Postgraduate Medicine.[1] In this study, Kochhar et al validated the BPHQ not for one, but for eleven languages spoken in India, which means that they embraced an effort equivalent of 11 validation studies. Their work and other validation studies have some noteworthy aspects to them. One such relevant aspect is the influence of the sample characteristics, such as the predominance of

References

in a particular sort of depression severity. For example, a study that has been developed with a sample consisting predominantly of subjects with four to six depressive symptoms (close to the cutoff of five symptoms for the diagnosis of MDD) would probably have a kappa statistic value lower than that which would be obtained if the study were developed with subjects that were predominantly in the extremes of the depressive symptomatology found in the criteria for MDD diagnosis (i.e., seven to nine symptoms or zero to one symptoms). In other words, in the first hypothetical sample, a difference of only one symptom between the instrument to be validated and the gold standard parameter could lead to a disagreement in the diagnosis in most of the cases. On the contrary, in the second hypothetical sample, even a disagreement of three symptoms between the instrument to be validated and the gold standard evaluation could still maintain both evaluations in agreement for a diagnosis of MDD for most patients. For example, patients with eight or nine depressive symptoms would still receive a diagnosis of MDD by the instrument to be validated, even if it had detected only five or six depressive symptoms. This would increase the number of agreements and consequently of the Kappa-value if the sample is consisted predominantly of patients with eight or nine depressive symptoms. An alternative approach to solve this problem has been described by Eaton et al., they proposed to consider the number of symptoms in disagreement instead of the simple disagreement for the diagnosis of depression.

Another aspect to be taken into consideration in the validation of an instrument in a new language is the elaboration of a back-translation. Such a procedure ensures the similarity of the new version with the original one. In the study, the kappa statistics were <0.5 for seven Indian languages. The authors had to improve the translations in these versions and rerun the validation process for them. The improvement in the kappa value with the translation adjustments was huge in the same version. For example, the Kappa for the Hindi version changed from 0.15 in the first run to 0.9 in the second run. It should be considered that performing a back translation could have made it possible to detect most of the translation biases and to save most of the extra work.

Two other points that we would like to bring up about validation studies are the establishment of the Kappa statistic as the parameter for validating an instrument and the cutoff for declaring the validity. The guidelines proposed by Landis and Koch, one of the most utilized, establishes a Kappa >0.6 as indication of substantial agreement and >0.8 as indication of an almost perfect agreement. Using the guidelines of Landis and Koch, the Kappa >0.5 used by the authors to declare a new Indian version of the BPHQ as valid is considered as being indicative of moderate agreement, which is adequate, in our point of view, considering the magnitude of their study. However, it should be mentioned that, as Landis and Koch commented, their cutoffs were completely arbitrary. It should also be mentioned that along with the Kappa statistics, the sensitivity and specificity parameters are also relevant in the validation process. For example, the Malayalam version of the BPHQ had a Kappa > 0.5 and this was a valid version. However, one should consider that although this version had an excellent specificity (0.96), its sensitivity was only 0.48. Consequently, this version is excellent for selecting patients such that one will be quite confident that they really have MDD when the instrument indicates that they do. However, it will fail to detect 52% of MDD cases.

In summary, validation studies should be encouraged and their interpretation and utility is better evaluated by taking into consideration the kappa statistics and other parameters such as sample characteristics, specificity and, sensitivity.

Fraguas R, Henriquez SG
Department and Institute of Psychiatry, Clinics Hospital, University of Sao Paulo School of Medicine, Sao Paulo, Brazil.
E-mail: rfraguas@hcnet.usp.br

References