Reliability and Prevalence in the DSM-5 Field Trials

January 12, 2012

Given the many changes proposed for the DSM-5, test-retest reliability was a major goal for the DSM-5 field trials. That is, would different clinicians, doing independent evaluations of the same patient less than 2 weeks apart, come to the same diagnostic conclusion? The DSM-5 field trials were designed to test this question of diagnostic reliability in conditions that approximated routine clinical diagnostic practice. Among the features of this design were:

- Patients were selected randomly with few exclusion criteria;
- Participating clinicians were fully qualified to make diagnoses, but not selected on the basis any special expertise in the disorders being evaluated;
- The entire DSM-5 diagnostic system was applied in each diagnostic evaluation, rather than focusing on one diagnosis at a time;
- Standardized diagnostic interviews, rarely used in routine clinical practice, were also not used in the field trials; clinicians were instructed to make their diagnoses based on their usual practices.

This emphasis on routine diagnostic practice departs from past field trials, and represents a new approach to the study of diagnostic reliability. Compared to the DSM-5 field trials, studies that have used expert raters, highly selected patients, or standardized diagnostic interviews can be expected to produce higher reliability for the tested diagnoses. Thus, it is important to provide realistic expectations for the reliability results from the DSM-5 field trials, and to use caution in making comparisons with past reliability studies. In addition, although the field trials were not designed to provide a detailed examination of DSM-IV diagnostic prevalence in comparison with DSM-5 prevalence for all disorders, information on the rates of DSM-IV diagnoses in each of the clinical settings was collected, as well as prevalence rates for the newly assessed DSM-5 diagnoses. There does not appear that there is a substantial difference in rates between DSM-IV and DSM-5, and in fact rates of DSM-5 diagnoses appear to be lower on average. This information will be published shortly.

Reliability is measured with the kappa statistic, which indicates the level of agreement between the clinicians making the diagnosis. Kappa ranges from 0 (no agreement between diagnosticians) to 1 (perfect agreement). There is no single agreed-upon standard for evaluating scores between 0 and 1. Studies in the rest of medicine were instructive in developing reasonable interpretations for the DSM-5 field trials before the results were available. Kramer et al. state, "for interrater reliability, in which two independent clinicians view, for example, the same X-ray or interview, one occasionally sees kappa values between 0.6 and 0.8, but the more common range is between 0.4 and 0.6 (1,2). For instance, in evaluating coronary angiograms, Detre et al. (3) reported that 'the level of observer agreement for most angiographic items (of 15 evaluated) [was] found to be approximately midway between chance expectation and 100% agreement' (i.e., kappa around 0.5). Test-retest studies are less frequent: the diagnosis of anemia based on conjunctival inspection was associated with kappa values between 0.36 and 0.60 (4), and the diagnosis of skin and soft-tissue infections was associated with kappa values between 0.39 and 0.43 (5). The test-retest reliability of various findings of bimanual pelvic examinations was associated with kappa values from 0.07 to 0.26 (6). From these results, to see a kappa for a DSM-5 diagnosis above 0.8 would be almost miraculous; to see κ between 0.6 and 0.8 would be cause for celebration. A realistic goal is kappa between 0.4 and 0.6, while κ between 0.2 and 0.4 would be acceptable…" (7)

No one is really satisfied with a kappa between .2 and .4, even though standard medical diagnoses do fall in this range. When a DSM-5 categorical diagnosis falls in this range, it will be seriously be reconsidered by the appropriate Work Group. However, when a disorder is rare, or the signs or symptoms are inconsistently expressed by patients with the disorder, it may be that in absence of a

biological test or repeated measures over time, a kappa between .2 and .4 may be as good as can be done using only a single clinical interview. Moreover, the fact that the usual DSM-5 diagnosis is categorical also places limits on its reliability. Dimensional diagnoses, included for the first time in DSM-5, are likely to be more reliable since they have the capacity of picking up more of the clinically important individual differences among patients.

Finally, Kraemer et al. note that "the Lancet (8) once described the evaluation of medical diagnostic tests as 'the backwoods of medical research,' pointing out that many books and articles have been written on the methods of evaluation of medical treatments, but little attention has been paid to the evaluation of the quality of diagnoses. Only recently has there been attention to standards for assessing diagnostic quality (9-11). Yet the impact of diagnostic quality on the quality and costs of patient care is great. Many medical diagnoses go into common use without any evaluation, and many believe that the rates of reliability and validity of diagnoses in other areas of medicine are much higher than they are. Indeed, psychiatry is the exception in that we have paid considerable attention to the reliability of our diagnoses. It is important that our expectations of DSM-5 diagnoses be viewed in the context of what is known about the reliability and validity of diagnoses throughout medicine and not be set unrealistically high, exceeding the standards that pertain to the rest of medicine." (7)

References

1. Koran LM: The reliability of clinical methods, data, and judgments (first of two parts). N Engl J Med 1975; 293:642–646
2. Koran LM: The reliability of clinical methods, data, and judgments (second of two parts). N Engl J Med 1975; 293:695–701
3. Detre KM, Wright E, Murphy ML, Takaro T: Observer agreement in evaluating coronary angiograms. Circulation 1975; 52:979–986
4. Wallace DE, McGreal GT, O'Toole G, Holloway P, Wallace M, McDermott EW, Blake J: The influence of experience and specialization on the reliability of a common clinical sign. Ann R Coll Surg Engl 2000; 82:336–338
5. Marin JR, Bilker W, Lautenbach E, Alpern ER: Reliability of clinical examinations for pediatric skin and softtissue infections. Pediatrics 2010; 126:925–930
6. Close RJ, Sachs CJ, Dyne PL: Reliability of bimanual pelvic examinations performed in emergency departments. West J Med 2001; 175:240–244
7. Kraemer HC, Kupfer DJ, Clarke DE, Narrow WE, Regier DA: DSM-5: How reliable is reliable enough? Am J Psychiatry 2012; 169:13-15 http://ajp.psychiatryonline.org/article.aspx?volume=169&page=13
8. The value of diagnostic tests (editorial). Lancet 1979; 1:809–810
9. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC , Lijmer JG, Standards for Reporting of Diagnostic Accuracy: The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003; 138:W1–W12
10. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC: Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Ann Intern Med 2003; 138:40–44
11. Meyer GJ: Guidelines for reporting information in studies of diagnostic test accuracy: the STARD initiative. J Pers Assess 2003; 81:191–193