DSM-5: How Reliable Is Reliable Enough?

DSM-5 is being developed for clinical decision making to provide the greatest possible assurance that those with a particular disorder will have it correctly identified (sensitivity) and that those without it will not have it mistakenly identified (specificity). Clinical diagnoses differ from diagnoses for other purposes: they are not necessarily sensitive enough for epidemiological studies or specific enough for basic and clinical research. We previously commented in these pages (1) on the need for field trials. Our purpose here is to set out realistic expectations concerning that assessment.

In setting those expectations, one contentious issue is whether it is important that the prevalence for diagnoses based on proposed criteria for DSM-5 match the prevalence for the corresponding DSM-IV diagnoses. However, to require that the prevalence remain unchanged is to require that any existing difference between true and DSM-IV prevalence be reproduced in DSM-5. Any effort to improve the sensitivity of DSM-IV criteria will result in higher prevalence rates, and any effort to improve the specificity of DSM-IV criteria will result in lower prevalence rates. Thus, there are no specific expectations about the prevalence of disorders in DSM-5. The evaluations primarily address reliability.

A DSM-5 field trial at a large clinical site is designed to draw a sample that is representative of the site’s patient population. These patients are to be evaluated independently by two clinicians, who are new to the patients, within an interval during which the presence or absence of the disorder is unlikely to have changed (i.e., between 4 hours and 2 weeks) to assess the test-retest reliability of the proposed diagnostic criteria. The clinicians are trained to use DSM-5 with training methods that would be available to any clinician. Reliability will be assessed using the intraclass kappa coefficient $k_I$ (2).

For a categorical diagnosis with prevalence $P$, among subjects with an initial positive diagnosis, the probability of a second positive diagnosis is $k_I + P(1-k_I)$, and among the remaining, it is $P(1-k_I)$. The difference between these probabilities is $k_I$ (3). Thus $k_I=0$ means that the first diagnosis has no predictive value for a second diagnosis, and $k_I=1$ means that the first diagnosis is perfectly predictive of a second diagnosis.

Reliability is essentially a signal-to-noise ratio indicator. In diagnosis, there are two major sources of “noise”: the inconsistency of expression of the diagnostic criteria by patients and the application of those criteria by the clinicians. It is all too easy to exaggerate reliability by removing some of that noise by design. Instead of a representative sample, as in DSM-5 field trials, one might select “case subjects” who are unequivocally symptomatic and “control subjects” who are unequivocally asymptomatic, omitting the ambiguous middle of the population for whom diagnostic errors are the most common and most costly. That approach would hide much of the patient-generated noise.

Moreover, there are three major types of reliability assessments that can be used depending on which sources of “noise” are permitted by design to affect diagnosis. Intrarater reliability requires that the same rater be asked to “blindly” review the same patient material two or more times. Noise related both to patients and to raters is removed. Interrater reliability requires that two or more different raters review the same patient material. Now the noise related to clinicians is included, but the noise related to

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.
patients is removed. Test-retest reliability requires that the same patients be observed separately by two or more raters within an interval during which the clinical conditions of the patients are unlikely to have changed. Now the noise related to both patients and to clinicians is included, as it would be in clinical practice. Consequently, for any diagnosis, intrarater reliability will be greater than interrater reliability, which will in turn be greater than test-retest reliability. It is test-retest reliability that reflects the effect of the diagnosis on clinical decision making and that is the focus of the DSM-5 field trials.

In addition, many reliability studies report “percentage agreement,” which substantially exaggerates reliability and fails to take into account agreement by chance. If diagnoses are randomly assigned a positive diagnosis with probability $P$, percentage agreement always exceeds 50% and approaches 100% when $P$ is either very large or very small. For example, when $P=0.95$, chance agreement would be 90%. The intraclass kappa is percentage agreement with chance agreement taken into account.

It is unrealistic to expect that the quality of psychiatric diagnoses can be much greater than that of diagnoses in other areas of medicine, where diagnoses are largely based on evidence that can be directly observed. Psychiatric diagnoses continue to be based on inferences derived from patient self-reports or observations of patient behavior. Nevertheless, we propose that the standard of evaluation of the test-retest reliability of DSM-5 be consistent with what is known about the reliability of diagnoses in other areas of medicine. Intrarater reliability is almost never assessed for psychiatric diagnosis because it is difficult to ensure blinding of two diagnoses by the same clinician viewing, for example, the same diagnostic interview. However, where intrarater reliability has been assessed for standard medical diagnostic procedures, it is common to see intrarater kappa values between 0.6 and 0.8 (4, 5), but there are exceptions (e.g., 0.54 for assessment of hand films for osteoarthrosis [4]).

Most medical reliability studies, including past DSM reliability studies, have been based on interrater reliability: two independent clinicians viewing, for example, the same X-ray or interview. While one occasionally sees interrater kappa values between 0.6 and 0.8, the more common range is between 0.4 and 0.6 (4, 5). For instance, in evaluating coronary angiograms, Detre et al. (6) 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., $k \approx 0.5$).

Examples in the medical literature of test-retest reliability are rare. The diagnosis of anemia based on conjunctival inspection was associated with kappa values between 0.36 and 0.60 (7), and the diagnosis of skin and soft-tissue infections was associated with kappa values between 0.39 and 0.43 (8). The test-retest reliability of various findings of bimanual pelvic examinations was associated with kappa values from 0.07 to 0.26 (9).

From these results, to see a $k_1$ for a DSM-5 diagnosis above 0.8 would be almost miraculous; to see $k_1$ between 0.6 and 0.8 would be cause for celebration. A realistic goal is $k_1$ between 0.4 and 0.6, while $k_1$ between 0.2 and 0.4 would be acceptable. We expect that the reliability (intraclass correlation coefficient) of DSM-5 dimensional measures will be larger, and we will aim for between 0.6 and 0.8 and accept between 0.4 and 0.6. The validity criteria in each case mirror those for reliability.

Generally, the lower kappa values are likely to occur with the rarer diagnoses. Thus, for a diagnosis with prevalence 0.05 and $k_1=0.2$, 24% of those with a positive first diagnosis and 4% of those with a negative first diagnosis will be positive on the second diagnosis (a risk ratio of 6.0). For a diagnosis with prevalence 0.5, our target would be closer to $k_1=0.5$, in which 75% with a positive first diagnosis and 24% with a negative first diagnosis would be positive on the second diagnosis (a risk ratio of 3.0).

The Lancet (10) 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 (11–13). 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.

References


HELENA CHMURA KRAEMER, PH.D.
DAVID J. KUPFER, M.D.
DIANA E. CLARKE, PH.D.
WILLIAM E. NARROW, M.D., M.P.H.
DARREL A. REGIER, M.D., M.P.H.

From Stanford University, Palo Alto, Calif.; University of Pittsburgh School of Medicine, Pittsburgh; and American Psychiatric Institute for Research and Education, American Psychiatric Association, Arlington, Va. Address correspondence to Dr. Kupfer (kupferdj@upmc.edu). Editorial accepted for publication May 2011 (doi: 10.1176/appi.ajp.2011.11010050).

The authors report no financial relationships with commercial interests.