Clinical Utility as a Criterion for Revising Psychiatric Diagnoses

Michael B. First, M.D.

Harold Alan Pincus, M.D.

John B. Levine, M.D., Ph.D.

Janet B.W. Williams, D.S.W.

Bedirhan Ustun, M.D.

Roger Peele, M.D.

Objective: Changes in DSM-IV were guided by empirical data that mostly focused on improving diagnostic validity and reliability. Although many changes were made explicitly to improve clinical utility, no formal effort was made to empirically determine actual improvements in clinical utility. The authors propose that future revisions of DSM empirically demonstrate improvement in clinical utility to clarify whether the advantages of changing the diagnostic criteria outweigh potential negative consequences.

Method: The authors provide a formal definition of clinical utility and then suggest that the merits of a proposed change to DSM be evaluated by considering 1) its impact on the use of the diagnostic system, 2) whether it enhances clinical decision making, and 3) whether it improves clinical outcome.

Results: Evaluating a change based on its impact on use considers both user acceptability and accuracy in application of the diagnostic criteria. User acceptability can be measured by surveying users' reactions, assessing user acceptability in a field trial setting, and measuring the effects on ease of use. Assessment of the correct application of diagnostic criteria entails comparing the clinician's diagnostic assessment to expert diagnostic assessment. Assessments of the impact on clinical decision making use methods developed for evaluating adherence to practice guidelines. Improvement in outcome entails measuring reduction in symptom severity or improvement in functioning or in documenting the prevention of a future negative outcome.

Conclusions: Empirical methods should be applied to the assessment of changes that purport to improve clinical utility in future revisions of DSM.

(Am J Psychiatry 2004; 161:946–954)

The DSM-IV revision process focused on empirically based evidence as the driving force of change (1). Empirical evidence that influenced DSM-IV decisions was obtained through three distinct but interactive stages: comprehensive reviews of the literature, reanalyses of previously collected data sets, and diagnosis-focused field trials (2). In most cases, data culled from empirical investigations were used with the express goal of improving the diagnostic validity of DSM. This effort to empirically validate psychiatric diagnoses was based on the seminal article by Robins and Guze (3), in which the authors outlined a method for establishing diagnostic validity. As described by Robins and Guze, the validity of these identified syndromes could be incrementally improved through increasingly precise clinical description, greater delineation of the syndromes from other disorders, laboratory studies, follow-up studies of outcome, and family studies. Over the years, investigators have proposed refining and expanding this list of external validators to include family history, demographic correlates, biological and psychological tests, environmental risk factors, concurrent symptoms (that are not part of the diagnostic criteria being assessed), treatment response, diagnostic stability, and course of illness (4). Although the paradigm of Robins and Guze has been criticized because after 30 years of research, it has not yet led to a nosology based on valid disease entities (5), the overall goal of iteratively improving the diagnostic validity of DSM based on empirical methods will continue to form the cornerstone of the DSM-V effort (6).

Complementary to ongoing efforts to improve the diagnostic validity of DSM, an equally important goal of the DSM revision process is to improve its clinical utility. The fundamental importance of clinical utility to DSM classification is exemplified by the statement in the introduction of DSM-IV-TR that its "highest priority has been to provide a helpful guide to clinical practice." But what exactly are the boundaries of the concept of clinical utility? The term appears frequently in the psychiatric literature; for example, a search of the PsycINFO literature database up to November 2002 found 229 articles with the term "clinical utility" in the title, the majority of which reported on the clinical utility of a test or treatment. Despite its common use, with the exception of a recent article by Kendell and Jablensky (7), we are not aware of any attempts to provide a formal definition of the concept. Although Kendell and Jablensky, in discussing the difference between diagnostic validity and utility, note that a diagnostic rubric has utility if "it can be shown to provide nontrivial information about prognosis and likely treatment outcomes, and testable propositions about biological and social correlates," we feel that this definition is overly narrow because it ignores some of DSM's more important clinical uses (e.g., communication of clinical information). Therefore, we propose the following definition of clinical utility, as it applies to the various components of DSM. Clinical utility is the extent to which DSM assists clinical decision makers in fulfilling the various clinical functions of a psychiatric classification system. These functions include assisting clinicians and other users with the following:

1. Conceptualizing diagnostic entities
2. Communicating clinical information to practitioners, patients and their families, and health care systems administrators
3. Using diagnostic categories and criteria sets in clinical practice (including for diagnostic interviewing and differential diagnosis)
4. Choosing effective interventions to improve clinical outcomes
5. Predicting future clinical management needs

Excluded from the concept of clinical utility are practical but nonclinical concerns such as the effect of a change on insurance reimbursements.

In fact, most of the changes in DSM-IV were made with the explicit goal of improving clinical utility. An example from the DSM-IV appendix, which provides an annotated listing of changes, includes "simplification" of the criteria sets for autistic disorder, conduct disorder, dementia, amnestic disorder, substance dependence, schizophrenia, generalized anxiety disorder, somatization disorder, and antisocial personality disorder. Another example is including new subtypes and disorders because of their "implications for treatment selection" (for example, "with atypical features," "with postpartum onset," "rapid cycling," and "bipolar II disorder"). In addition, the goal of improving clinical utility can be inferred from most of the other annotations, which include phrases such as "changed to reflect clinical use" and "allow for earlier case finding."

Even though a number of these proposed changes in criteria were studied in the DSM-IV field trials (8–12) or in data reanalyses (13), no formal effort was made to empirically examine whether these changes actually improved clinical utility. Instead, the field trials and data reanalyses primarily evaluated proposed criteria sets in terms of reliability, validity (using clinical diagnoses as the standard), and the extent to which the proposed criteria set identified different individuals as having the disorder. Purported improvements in clinical utility were simply assumed to be the case.

In this article, we propose that the same standards for empirically based documentation that have been applied to the determination of whether a change improves diagnostic validity also be applied to changes that aim to improve clinical utility. We specifically recommend that future field trials incorporate specific assessments that quantify the extent of the improvement in clinical utility. One might wonder why improvements in clinical utility need to be empirically documented. For example, it might appear intuitively obvious that replacing a criterion that requires five items from a list of 15 with one that requires only two items of six is an improvement in clinical utility. Determining improvement in clinical utility by using empirical methods would serve to clarify whether the advantages of changing the diagnostic criteria outweigh potential negative consequences. Such negative consequences include the burden on users to learn about the changes, the burden on instrument developers to incorporate the changes, and the difficulties imposed on researchers by complicating their ability to pool or compare data from studies using different versions of criteria (14).

The Relationship Between Diagnostic Validity and Clinical Utility

Although in this article we discuss diagnostic validity and clinical utility as if they were distinct concepts, there is, in fact, considerable debate about the overlap between these two constructs, largely because of the lack of widely accepted definitions of these constructs as they apply to psychiatric diagnoses. At one extreme are those that use the terms interchangeably. For example, Spitzer (15) defines diagnostic validity as "the extent [to which] the defining features of a disorder provide useful information not contained in the definition of the disorder." At the other extreme, Kendell and Jablensky (7) see validity and utility as completely distinct constructs, arguing that a disorder can be considered valid only if it can be demonstrated to have natural boundaries with other disorders.

Our position is to take the middle ground. Although we view validity and utility as separate constructs, there is considerable overlap between them. Diagnostic validity is a complex multifaceted construct that has historically been adapted from the field of psychological testing and includes a number of different types of validity. These include the following:

1. Face validity (i.e., whether the description of a category and its diagnostic criteria seem to accurately describe the disorder)
2. Descriptive validity (i.e., whether the features of a category are unique to that category relative to other mental disorders)
3. Predictive validity (i.e., the extent to which having a diagnosis predicts future clinical course, complications, and treatment response)
4. External or construct validity (i.e., the extent to which the diagnosis correlates with expected external validators, such as family history and neurobiological markers) (16, 17)

Since many of these elements of diagnostic validity are inherently useful in the care of patients (e.g., in predicting treatment response or course of illness), it is not surprising that improving diagnostic validity often improves clinical utility as well.

There are a number of scenarios in which there is no overlap between diagnostic validity and clinical utility, and it is conceivable that a change intended to improve the clinical utility of DSM might reduce diagnostic validity. For example, changes in the diagnostic criteria sets that are designed to make them easier to use (e.g., drastically shortening the number of diagnostic criteria) could reduce diagnostic validity (e.g., weaken their association with an external validator such as family history). One way to minimize this potential conflict is to ensure that simplified criteria sets identify the same patients as having the disorder, thus guaranteeing that diagnostic validity is maintained. For this reason, an important aspect of several of the DSM-IV field trials (e.g., for somatization disorder and antisocial personality disorder) was the determination of whether the simplified DSM-IV criteria sets diagnosed the same set of individuals as did the more complex DSM-III-R criteria sets. Although it is theoretically possible to make a change in DSM that would improve diagnostic validity without also improving its clinical utility (e.g., adding a genetic subtyping scheme that has absolutely no management implications), we are not aware of any such examples in psychiatry.

Developing an Empirical Method for Assessing Change in Clinical Utility

When evaluating whether a proposed change improves or detracts from the clinical utility of DSM, it is important to first identify the various components of clinical utility that might be affected by that change. Based on the definition of clinical utility that we proposed, we suggest that the merits of a proposed change to DSM can be evaluated by considering 1) its impact on the use of DSM, 2) whether it enhances clinical decision making, and 3) whether it improves clinical outcomes. These three components of clinical utility imperfectly resemble the classic "structure/process/outcome" framework for evaluating quality of care set forth by Donabedian (18). Each of these will be discussed in more detail.

Impact on Use

This aspect of clinical utility focuses on whether the diagnostic system is used *at all* by its intended end population (user acceptability) and whether it is used *correctly* (accuracy in the application of the diagnostic criteria). Achieving adequate levels of user acceptability is critically important since nonutilization of the diagnostic system eliminates any potential benefits that might result from DSM changes that target either clinical decision making or patient outcome. Similarly, the correct application of the DSM diagnostic criteria in clinical settings is also a prerequisite for assigning the proper psychiatric diagnosis. Selecting the wrong diagnosis from the get-go may subsequently lead to improper treatment selection and ultimately to poor patient outcomes.

Establishing Current Use

Before one can determine whether a proposed change to DSM positively affects its use, it is important to establish baseline information about how the current version of DSM is being used and to identify potential sources of poor user acceptability. Surprisingly, relatively little research has focused on the clinical use of DSM. Two methods have been used: 1) surveys of clinicians' and researchers' self-reported use of—and attitudes toward—various editions of the DSM and ICD classifications and 2) direct measurement of use through examination of recorded chart diagnoses. Although a number of surveys have been conducted over the past 20 years (19–26), their focus has been almost exclusively on general system-wide issues rather than on discerning attitudes toward specific diagnostic categories, Such restrictions have limited conclusions about the user acceptability of specific diagnostic categories.

One exception to this approach in survey questions concerns the DSM multiaxial system. A major impetus for overhauling DSM-III-R axis IV was the concern that it was "used infrequently in clinical settings and research studies" (27). As part of the literature review of articles examining the use, reliability, and validity of axis IV, Skodol examined the surveys of DSM-III and DSM-III-R users that included questions regarding their use of axis IV. For example, in the 1986 survey of practicing psychiatrists and 1984 graduates of residency programs (21), only 44% of the psychiatrists and 56% of the residents found axis IV to be useful. Partly based on the surveyed underuse of axis IV, the DSM-IV work group ultimately recommended that axis IV be replaced by a psychosocial and environmental problem list. It should be noted that although the executive summary of changes in the *DSM-IV Sourcebook* (28) noted that the psychosocial and environmental checklist was added because of "its simplicity and the fact that it allows clinicians to note the specific problems of concern," no field trials were conducted to empirically determine whether this assertion was, in fact, true.

An alternative to assessing use by surveying clinicians is to measure it by reviewing clinical records for evidence of use. This method formed the basis of the recommendation by the Psychiatric Systems Interface Disorders Work Group to delete three of the subtypes for adjustment disorder (i.e., adjustment disorder with physical complaints, adjustment disorder with withdrawal, and adjustment disorder with work/academic inhibition) because they were "rarely used" (29). As part of the MacArthur data reanalysis conducted by Strain and colleagues (30), an analysis of the frequency distribution of DSM-III subtypes of adjustment disorder from a National Institute of Mental Health (NIMH) sample (6,654 patients diagnosed with adjustment disorder) and a database of patients seen at the Adult Psychiatry Crisis Clinic at Western Psychiatric Institute and Clinic (2,916 diagnosed with adjustment disorder) indicated that "with withdrawal" and "with work/academic inhibition" were almost never used. These findings led the work group to recommend deletion of these subtypes for simplification.

An important limitation in measuring use by looking at chart diagnoses is that they may bear little relationship to actual use of DSM or to the potential clinical utility of a diagnosis or subtype. Some clinicians may record a particular chart diagnosis to receive payment for the treatment they provide, to minimize stigmatization, or because it fits the "gestalt" of the patient, rather than because the diagnostic criteria were carefully evaluated and determined to have been met. More extensive evaluations of diagnostic procedures may be needed to draw accurate conclusions about DSM use. Such evaluations might entail requiring a much more extensive documentation of whether the diagnostic criteria have been met or conducting post hoc "debriefing sessions" in which the clinician's diagnostic decision making is carefully examined by using cognitive analysis techniques (31).

Assessing User Acceptability of Proposed Changes

User acceptability of a proposed change can be divided into two components, each of which can be assessed separately: 1) confidence in its diagnostic validity and 2) its ease of use. For example, the user acceptability of the DSM-IV proposal to introduce the diagnosis of Asperger's disorder to the section on pervasive developmental disorders depends both on whether the user believes that the DSM-IV diagnosis is a valid construct and whether the criteria set is easy enough to use. However, any given clinician's perception that a proposed diagnostic construct is valid will depend on a number of personal factors, such as familiarity with the scientific literature (relevant for deciding whether a proposed change has external validity), clinical experience (relevant for deciding about predictive validity and face validity), and practice setting (since case mix can affect a clinician's perception of what is descriptively valid). In light of these limitations, we are certainly not arguing that diagnostic changes should be based primarily on their popularity with clinicians. Nonetheless, to ensure that proposed changes are in line with overall clinical sensibilities, some assessment of clinicians' attitudes about the credibility of proposed changes is advisable, as long as the inherently subjective nature of such assessments is acknowledged. Furthermore, surveys should attempt to explore the reasons behind clinicians' judgments regarding the acceptability or unacceptability (use or nonuse) of a particular disorder.

The ease of use or "user friendliness" of a proposed diagnostic change is also crucial in determining its ultimate user acceptability. This parameter covers the practical issues of applying the diagnostic system and includes the length of time it takes to assess a particular criteria set, the number and complexity of criteria that affect the likelihood that the user will be able to recall from memory a criteria set (thus obviating the need to have the DSM manual always on hand), and the ease in locating the appropriate disorder in the classification.

User acceptability of a proposed change can be measured in a number of different ways, including the following:

1. Surveying users' reactions to a presentation of the proposed changes
2. Assessing user acceptability in the context of actual use (e.g., in a field trial setting)
3. Measuring the proposed changes' effects on ease of use (e.g., by timing the duration of assessment procedures)

Surveys of users' reactions to proposed changes can help in determining the likelihood that such changes will be accepted by the DSM user community at large. Using the survey methods, the DSM-IV Child Disorders Work Group queried 460 child psychiatrists about their reactions to specific proposed revisions for DSM-IV using a questionnaire composed of 52 items that was administered to attendees of the 1989 meeting of the American Academy of Child Psychiatry (24). Proposed changes were described, and the respondents were asked whether they supported the proposals. An example of such an item included in the survey was "the work group plans to recommend that reactive attachment disorder be differentiated into socially withdrawn and socially indiscriminate subtypes." In some cases, the respondents were asked specific questions that might help provide the work group with particular use information for its deliberations. For example, because the questionable empirical basis of identity disorder resulted in a proposal that the category be deleted, the DSM-IV Childhood Disorders Work Group was interested in how many practitioners actually use that diagnosis. Of note, even though the survey indicated that 80% reported that they use it for adolescents and 45% reported using it for children, the lack of any research data regarding its empirical validity took precedence over the fact that it was used clinically and identity disorder was ultimately removed from the mental disorders section of DSM-IV.

Field trial methods usually ask users to apply diagnostic criteria to case vignettes or in actual clinical settings. The international field trials of the ICD-10 clinical guidelines (32) and diagnostic criteria for research (33) both measured user acceptability by applying a draft of the ICD-10 clinical guidelines and diagnostic criteria for research to written case histories and to actual patients. In addition to recording the diagnosis on a rating form, clinicians were asked to rate on a 5-point scale how well the diagnosis provided a “good fit” to the patient’s clinical picture, how confident they were in making the diagnosis using the clinical guidelines, and how easy they were to use. These field trials were useful in identifying inconsistencies and ambiguities in the clinical guidelines and diagnostic criteria for research, leading to a number of changes and improvements.

One problem with measuring users’ attitudes about the acceptability of diagnostic changes is the lack of any available standardized instruments. To date, existing questionnaires have been developed ad hoc for particular surveys or field trials with no attempts having been made to determine such basic psychometric properties as internal consistency of the responses or test-retest reliability. We recommend that efforts be made to develop standardized assessments so that results can be compared across different studies to determine relative levels of user acceptability.

User acceptability can also be assessed indirectly by measuring the effect of a proposed change on the ease of use in an experimental setting. The DSM-III-R criteria set for generalized anxiety disorder was revised in DSM-IV, replacing the 18-item criterion D (which listed anxiety symptoms often present) with a corresponding six-item criterion C that is purported to be “simpler, more reliable, and more coherent” based on the results of a data reanalysis (13). The data reanalysis evaluated whether the shorter criteria set improved reliability (which it did), but no attempt was made to empirically evaluate whether in fact the modified criteria for generalized anxiety disorder were easier to use. To measure increased ease of use in an experimental setting, the two criteria sets could be compared in terms of time required for assessment. In addition, ease of use could also be measured by determining which of the criteria sets was easier for the clinician to recall from memory.

Assessing the Accuracy of Application of Diagnostic Criteria

In addition to potentially improving user acceptability, proposed changes that make the criteria sets easier to use can improve users’ ability to correctly apply the diagnostic criteria in practice. These include proposals to simplify criteria sets or diagnostic algorithms, clarifications in the wording of an ambiguously written diagnostic criterion, changes in criteria to make them easier to assess, and modifications in exclusion criteria to more clearly delineate the differential diagnosis. For example, the diagnostic criteria for schizoaffective disorder in DSM-III-R were ambiguous in that it was unclear whether the user was supposed to focus on the current episode of illness or on the lifetime pattern when determining the temporal relationship of mood and psychotic symptoms (34). For DSM-IV, the wording of the diagnostic criteria was changed to clarify that all of these criteria concern an uninterrupted period of illness rather than the lifetime pattern of symptoms, with the expectation that clinicians using the revised diagnostic criteria will be more likely to apply them correctly from a procedural standpoint. It should be noted that the goal of these wording changes was not to improve the diagnostic validity of the schizoaffective disorder construct per se. That is, these changes were not intended to indicate that a definition of schizoaffective disorder that focuses on the relationship of mood and psychotic symptoms within a continuous episode is more valid than a definition that concentrates on the lifetime pattern of mood and psychotic symptoms. Instead, these changes to the criteria set were made to ensure that the operationalization of the schizoaffective disorder construct was consistent across different diagnostic criteria.

Assessment of whether clinicians are “correctly” applying the diagnostic criteria usually entails comparing the clinician’s application of the diagnostic criteria with an expert’s. This method has been used in a number of studies to confirm diagnostic accuracy (e.g., reference 35, in which the accuracy of the clinician’s diagnosis is measured by virtue of its agreement with the best estimate diagnosis by using a structured interview that promotes rigorous application of the diagnostic criteria). Whether the proposed change in the diagnostic criteria for schizoaffective disorder actually improves diagnostic accuracy can be assessed by comparing clinicians’ ability to agree with a Longitudinal Expert All Data (LEAD) diagnosis (36) using the DSM-III-R versus DSM-IV diagnostic criteria.

Impact on Clinical Decision Making

Clinicians do not use DSM simply to assign a diagnostic label (or labels) to a patient’s clinical presentation. Rather, the DSM diagnosis informs a variety of clinical decisions by both the clinician and the patient. For the clinician, these decisions include selection of a particular setting for treatment (e.g., inpatient versus outpatient), mode of treatment (e.g., somatic and/or psychosocial), and expected duration of treatment (e.g., brief versus long-term). Patients may also alter their behavior once their psychiatric diagnosis and its implications are explained to them (e.g., avoiding known triggers of acute episodes of illness). Most changes that entail the addition of a new disorder or new subtype are intended to highlight a homogeneous (and previously unidentified) group of patients that require special clinical attention. For example, the addition in DSM-IV of the rapid cycling subtype for bipolar disorder was intended to alert clinicians, patients, and their families to this important subgroup of bipolar patients and to encourage clinicians, patients, and their families to follow certain management guidelines (e.g., for clinicians, measurement of thyroid hormone, use of valproate in lieu of lithium, and exercise of extra caution in the use of antidepressants, and for patients, the avoidance of potential triggers such as substance use or sleep deprivation).

Assessments of whether clinicians are making the appropriate clinical decisions can be done by using methods developed for evaluating the degree of adherence to practice guidelines. Such methods involve, first, the delineation of "quality indicators," which consist of specific accessible data elements that indicate a reasonable likelihood of concordance with a particular guideline (37, 38). These indicators usually consist of a denominator specifying the population to which the guideline applies (e.g., all individuals with an insurance claim for bipolar disorder) and a numerator suggesting conformance with the guidelines (e.g., all individuals in the denominator who have been given a prescription for a mood stabilizer). Data can be collected from clinician or patient surveys, systematic chart abstraction, or secondary analyses of electronic medical records or claims data.

There are a number of ways to design studies to examine the impact of a proposed change on clinical decision making. The ideal design would document an increased incidence of the desired clinical decision in a randomized controlled trial in which patients are randomly assigned to one of two groups: patients diagnosed according to the existing diagnostic classification and patients diagnosed according to the proposed changes. This design has the advantage of controlling for other factors that might affect the clinical decision-making process. Other designs that are less ideal but still useful could examine claims forms, encounter data, or patient charts to document increased use of the proposed diagnostic specifier in association with the desired clinical decision (e.g., use of valproate associated with use of the rapid-cycling specifier).

For example, the "atypical features" specifier was added to DSM-IV because it identified a subgroup of depressed patients that respond less well to tricyclic antidepressants (39). In order to determine whether this addition to DSM-IV is clinically useful (in regard to the effect of the change on clinical decision making), one could conduct a randomized controlled trial that would examine the rates of use of tricyclic antidepressants in groups of depressed patients with a mix of subtypes who were randomly assigned to groups according to whether they are diagnosed according to DSM-III-R or DSM-IV. Improvement in clinical utility could be demonstrated if the rates of use of tricyclic antidepressants in patients diagnosed according to DSM-IV criteria (some of whom would have been identified as having atypical features and thus not given tricyclic antidepressants) were shown to be lower compared to patients diagnosed according to DSM-III-R criteria (none of whom would have been labeled as atypical). A study examining chart data for an association between the atypical features specifier and less frequent use of tricyclic antidepressants would be problematic since it would not be able to account for other reasons for lower use of tricyclic antidepressants (e.g., the potential for lethality in overdose and increased rates of problematic side effects).

Improvement in Clinical Outcome

The "holy grail" of clinical utility is the positive effect of a change in the diagnostic system on outcome. Improvement in clinical outcome can be assessed by measuring reduction in symptom severity, measuring improvement in functioning, or documenting the prevention of a future negative outcome (e.g., reduction in relapse rates over a period of time). Given the number and complexity of the determinants of outcome, it has traditionally not been possible to empirically demonstrate that a change in diagnostic practices results in an improvement in outcome because of the very large group sizes that would be required to demonstrate a difference among groups. However, in response to severe limitations in the utility and generalizability of typical clinical trials, the NIMH has initiated several large-scale, public-health-oriented clinical trials that have a more naturalistic design (40). In such trials, diagnostic algorithms are typically tested in a wide variety of clinical settings, comparing a "treatment-as-usual" arm with various permutations of the diagnostic algorithm. Trials of this type may potentially provide an opportunity to relate clinical outcome to diagnostic innovations (e.g., treatment-oriented specifiers such as rapid cycling, seasonal pattern, and atypical features). Studies that employ treatment algorithms that incorporate existing or proposed DSM diagnostic considerations in their decision points (i.e., if the patient has a particular DSM subtype, then institute a particular treatment) may allow for an examination of the effects of a diagnostic change on outcome by allowing comparison with a treatment-(and diagnosis)-as-usual arm.

Examples of Applying Elements of Clinical Utility

To illustrate how the impact of a particular change on clinical utility might be measured, we will use as examples two changes proposed for, and ultimately included in, DSM-IV.

1. Simplification of the Diagnostic Criteria for Somatization Disorder

One of the explicit goals of the proposed change in the DSM-IV criteria set for somatization disorder was to improve its clinical utility (11). As a result of a data reanalysis of 500 patients with somatization disorder, a new criteria set was proposed, replacing the DSM-III-R requirement for at least 13 from an exhaustive list of 35 somatoform symptoms with a requirement for a pattern of symptoms drawn from different types (i.e., at least four pain symptoms, two gastrointestinal, one sexual, and one pseudoneurological). Although the DSM-IV field trial documented that the revised criteria set defined the same group of patients as did the DSM-III-R criteria set, no empirical efforts were made to actually document that the revised criteria set was in fact easier to use. Presumably, the opinion of the work group that the revised criteria sets would be substantially easier to use was sufficient.

How could the clinical utility aspects of the proposed criteria set have been measured? One of the justifications for making this change was the belief that clinicians found the DSM-III-R criteria set too difficult to use and thus were unlikely to apply the somatization disorder criteria set in their practice. (It should be noted that this justification was *not* based on any empirical data that showed that the clinicians harbored negative opinions about the somatization disorder criteria or that they were in fact not using these criteria in practice.) Thus, the primary component of clinical utility that was targeted was the clinicians' acceptability of the somatization disorder criteria. This could be measured as part of a field trial by first assessing (by means of a questionnaire) whether the clinicians in the trial thought that the proposed criteria were easier to use and whether the criteria set captured the "clinical essence" of somatization disorder. In addition, ease of use could be objectively measured by comparing the actual time it takes to do the evaluation using the DSM-III-R and DSM-IV criteria sets and/or to compare the ability of users to remember the diagnostic algorithm. Finally, the ability of clinicians to accurately apply the proposed diagnostic criteria could be assessed by comparing the agreement of clinicians' diagnoses with an expert standard for each of the two criteria sets.

2. Addition of the Bipolar II Disorder Category to DSM-IV

Bipolar II disorder is defined as a pattern of major depressive episodes and hypomanic episodes. Before its inclusion in DSM-IV, these patients would usually have been diagnosed as having major depressive disorder, although, potentially, patients with accompanying episodes of severe hypomania could have been diagnosed as having bipolar disorder not otherwise specified. Three reasons were cited (39) for the addition of bipolar II disorder in DSM-IV:

1. The disorder retained its course over a 5-year follow-up (i.e., only 10% of patients developed manic episodes).
2. Family studies of probands with bipolar II disorder demonstrated higher rates of bipolar II disorder (and bipolar I disorder) compared with community rates.
3. There was a "clinical and research need to identify optimal treatment strategies for this somewhat distinct condition" (39, p. 1020).

Although the first two goals are more closely related to diagnostic validity, the third goal is more directly related to clinical utility.

The various components of clinical utility just described can be examined. With regard to user acceptability, users' perception of its diagnostic validity can be assessed by using surveys administered to representative samples of clinicians (e.g., "Do you think the addition of bipolar II disorder would identify a specific subgroup with important clinical implications?" "Does this criteria set capture your clinical impression of the disorder?"). Ease of use and accuracy of use can be measured by conducting a field trial of clinicians who would be asked to apply the diagnostic criteria for bipolar II in an appropriate setting (e.g., a mood disorders clinic). Ease of use can be measured by surveying clinicians to determine the extent to which they think that the criteria (especially the diagnostic criteria for a hypomanic episode) were easier to use. Correct use can be measured by comparing the clinician's diagnosis to a diagnostic assessment conducted by an expert on a subgroup of the patients.

The effect of bipolar II on clinical decision making (e.g., treatment selection) can be measured by determining adherence to treatment guidelines for bipolar II disorder. Although evidence-based treatment guidelines for bipolar II disorder have not yet been firmly established, the 1996 expert consensus guidelines for the treatment of bipolar disorder generally recommend the addition of mood stabilizers to antidepressant regimens (41). Thus, one indication of whether decision making on the clinician side is improved might be whether the use of mood stabilizers is associated with being given a diagnosis of bipolar II disorder in this population. To determine whether a patient's decision making is modified, one could document whether there is an association between patients receiving this diagnosis and evidence that they have modified their behavior in order to reduce the risk of recurrence (e.g., adherence to maintenance medication regimens). Alternatively, improvement in clinical decision making could be documented by means of a randomized, controlled trial in which patients were randomly assigned to two groups of clinicians, with one group diagnosed according to DSM-III-R criteria and the other according to DSM-IV criteria.

The ultimate measure of clinical utility for this population would be an improvement in outcome. An outcome study could be designed to compare clinical outcomes between groups of patients with both depression and some hypomanic symptoms diagnosed according to DSM-III-R versus patients diagnosed according to DSM-IV. Targets for assessment might include demonstrating a better clinical course (i.e., fewer depressive and hypomanic episodes, that the depressive episodes are of shorter duration, and that the patients are less likely to develop full-blown manic episodes) and better social and occupational functioning.

Conclusions

In comparison to DSMs of the past, the DSM-IV revision process moved from an almost sole reliance on expert consensus to a greater reliance on reviews of published literature, data set reanalyses, and field trials focused on diagnostic issues. Although many of the changes in DSM-IV aimed to improve its clinical utility, the focus of the DSM-

IV literature reviews and empirical studies were almost exclusively on diagnostic validity. Given the important role that DSM serves in facilitating clinical practice, an equally crucial target for evaluating the advantages and disadvantages of a particular change is its effect on clinical utility. Improvements in clinical utility can be measured in terms of 1) their impact on the use of DSM, 2) their enhancement of clinical decision making, and 3) whether they lead to improvement in clinical outcomes. Given the current lack of standardized instruments to measure user acceptability and ease of use, efforts should be made to develop psychometrically sound assessment tools that can be used in future studies. Only by applying empirical methods for assessing improvement in clinical utility can it be demonstrated that purported improvements in clinical utility are, in fact, achievable and that the magnitude of these improvements in clinical utility justifies the inevitable disruption that comes with changing the diagnostic criteria.

Received Jan. 22, 2003; revision received April 2, 2003; accepted July 15, 2003. From the Department of Psychiatry, Columbia University, New York; Western Psychiatric Institute and Clinic, University of Pittsburgh, Pittsburgh; Rand Corporation, Pittsburgh; the Department of Psychiatry, Harvard Medical School, Boston; Classification, Assessment Surveys, and Terminology, World Health Organization, Geneva, Switzerland, and the Psychiatry Department, George Washington University, Washington, DC. Address reprint requests to Dr. First, New York State Psychiatric Institute, 1051 Riverside Dr., Unit 60, New York, NY 10032; mbf2@columbia.edu (e-mail).

Dr. Pincus is supported in part by the NIMH-funded Mental Health Intervention Research Center on Mood and Anxiety Disorders, the University of Pittsburgh, and the Robert Wood Johnson Foundation.

The authors comprise an ad hoc subcommittee of the American Psychiatric Association's Committee on Psychiatric Diagnosis and Assessment. This subcommittee was given the charge of considering the role of clinical utility and approaches to developing an associated evidence base in future revisions of DSM. Although the views expressed in this article have been endorsed by the Committee on Psychiatric Diagnosis and Assessment, they do not necessarily represent the policies and opinions of the American Psychiatric Association.

References

1. Frances A, First M, Pincus H: DSM-IV Guidebook. Washington, DC, American Psychiatric Press, 1995
2. Widiger T, Frances A, Pincus H, Davis W, First M: Toward an empirical classification for the DSM-IV. J Abnorm Psychol 1991; 100:280–288
3. Robins E, Guze SB: Establishment of diagnostic validity in psychiatric illness: its application to schizophrenia. Am J Psychiatry 1970; 126:983–987
4. Kendler K: Toward a scientific psychiatric nosology: strengths and limitations. Arch Gen Psychiatry 1990; 47:969–973
5. Hyman S: Neuroscience, genetics, and the future of psychiatric diagnosis. Psychopathology 2002; 35:139–144
6. Kupfer D, First M, Regier D (eds): A Research Agenda for DSM-V. Washington, DC, American Psychiatric Press, 2002
7. Kendell R, Jablensky A: Distinguishing between the validity and utility of psychiatric diagnoses. Am J Psychiatry 2003; 160:4–12
8. Volkmar FR, Klin A, Siegel B, Szatmari P, Lord C, Campbell M, Freeman BJ, Cicchetti DV, Rutter M, Kline W, et al: Field trial for autistic disorder in DSM-IV. Am J Psychiatry 1994; 151:1361–1367
9. Lahey BB, Applegate B, Barkley RA, Garfinkel B, McBurnett K, Kerdyk L, Greenhill L, Hynd GW, Frick PJ, Newcorn J, et al: DSM-IV field trials for oppositional defiant disorder and conduct disorder in children and adolescents, in DSM-IV Sourcebook, vol 4. Edited by Widiger TA, Frances AJ, Pincus HA, Ross R, First MB, Davis W, Kline M. Washington, DC, American Psychiatric Association, 1998, pp 661–686
10. Flaum M, Amador X, Gorman J, Bracha HA, Edell W, McGlashan T, Pandurangi A, Kendler KS, Robinson D, Lieberman J, et al: DSM-IV field trials for schizophrenia and other psychotic disorders. Ibid, pp 687–716
11. Yutzy SH, Cloninger CR, Guze SB, Pribor EF, Martin RL, Kathol RG, Smith GR, Strain JJ: DSM-IV field trial: testing a new proposal for somatization disorder. Am J Psychiatry 1995; 152:97–101
12. Widiger TA, Cadoret R, Hare R, Robins L, Rutherford M, Zanarini M, Alterman A, Apple M, Corbitt E, Forth A, Hart S, Kultermann J, Woody G, Frances A: DSM-IV antisocial personality disorder field trial. J Abnorm Psychol 1996; 105:3–16
13. DiNardo P: Generalized anxiety disorder, in DSM-IV Sourcebook, vol 4. Edited by Widiger TA, Frances AJ, Pincus HA, Ross R, First MB, Davis W, Kline M. Washington, DC, American Psychiatric Association, 1998, pp 259–266
14. Rounsaville B, Alarcon R, Andrews G, Jackson J, Kendell R, Kendler K: Basic nomenclature issues for DSM-V, in A Research Agenda for DSM-V. Edited by Kupfer D, First M, Regier D. Washington, DC, American Psychiatric Press, 2002, pp 1–30
15. Spitzer RL: Values and assumptions in the development of DSM-III and DSM-III-R: an insider's perspective and a belated response to Sadler, Hulhus and Agich's "On Values in Recent American Classification." J Nerv Ment Dis 2001; 189:351–359
16. Spitzer R, Williams J: Classification of mental disorders, in Comprehensive Textbook of Psychiatry, 4th ed. Edited by Kaplan HI, Sadock BJ. Baltimore, Williams & Wilkins, 1985, pp 591–612
17. Blacker D, Endicott J: Psychometric properties: concepts of reliability and validity, in Handbook of Psychiatric Measures. Washington, DC, American Psychiatric Association, 2000, pp 7–14
18. Donabedian A: The quality of medical care, in Medicine in a Changing Society. Edited by Corey L, Saltman SE, Epstein MF. St Louis, CV Mosby, 1972, pp 83–101
19. Junek RW: The DSM-III in Canada: a survey. Can J Psychiatry 1983; 28:182–187
20. Mezzich AC, Mezzich JE: Perceived suitability and usefulness of DSM-III vs DSM-II in child psychopathology. J Am Acad Child Psychiatry 1985; 24:281–285
21. Jampala VC, Sierles FS, Taylor MA: Consumers' views of DSM-III: attitudes and practices of US psychiatrists and 1984 graduating psychiatric residents. Am J Psychiatry 1986; 143:148–153
22. Smith D, Kraft WA: Attitudes of psychiatrists toward diagnostic options and issues. Psychiatry 1989; 52:66–78
23. Maser JD, Kaelber C, Weise RE: International use and attitudes toward DSM-III and DSM-III-R: growing consensus in psychiatric classification. J Abnorm Psychol 1991; 100:271–279
24. Setterberg SR, Ernst M, Rao U, Campbell M, Carlson GA, Shaffer D, Staghezza BM: Child psychiatrists' views of DSM-III-R: a survey of usage and opinions. J Am Acad Child Adolesc Psychiatry 1991; 30:652–658
25. Jampala VC, Zimmerman M, Sierles FS, Taylor MA: Consumers' attitudes toward DSM-III and DSM-III-R: a 1989 survey of psychiatric educators, researchers, practitioners, and senior residents. Compr Psychiatry 1992; 33:180–185
26. Mezzich JE: International surveys on the use of ICD-10 and related diagnostic systems. Psychopathology 2002; 35:72–75
27. Skodol A: Axis IV, in DSM-IV Sourcebook, vol 3. Edited by Widiger TA, Frances AJ, Pincus HA, Ross R, First MB, Davis W. Wash-

ington, DC, American Psychiatric Association, 1997, pp 409–422

28. Williams JBW: DSM-IV multiaxial system: final overview, in DSM-IV Sourcebook, vol 4. Edited by Widiger TA, Frances AJ, Pincus HA, Ross R, First MB, Davis W, Kline M. Washington, DC, American Psychiatric Association, 1998, pp 939–946
29. Hales RE: DSM-IV Psychiatric System Interface Disorders (PSID) Work Group: final overview. Ibid, pp 1077–1086
30. Strain JJ, Newcorn J, Mezzich J, Kirisci L: Adjustment disorder: the MacArthur reanalysis. Ibid, pp 403–426
31. Patel VL, Arocha JF, Diermeier M, Greenes RA, Shortliffe EH: Methods of cognitive analysis to support the design and evaluation of biomedical systems: the case of clinical practice guidelines. J Biomed Inform 2001; 34:52–66
32. Sartorius N, Kaelber CT, Cooper JE, Roper MT, Rae DS, Gulbinat W, Ustun TB, Regier DA: Progress toward achieving a common language in psychiatry: results from the field trial of the clinical guidelines accompanying the WHO classification of mental and behavioral disorders in ICD-10. Arch Gen Psychiatry 1993; 50: 115–124
33. Sartorius N, Üstün TB, Korten A, Cooper JE, van Drimmelen J: Progress toward achieving a common language in psychiatry, II: results from the international field trials of the ICD-10 diagnostic criteria for research for mental and behavioral disorders. Am J Psychiatry 1995; 152:1427–1437
34. Flaum M, Andreasen NC, Widiger TA: Schizophrenia and other psychotic disorders in DSM-IV: final overview, in DSM-IV Sourcebook, vol 4. Edited by Widiger TA, Frances AJ, Pincus HA, Ross R, First MB, Davis W, Kline M. Washington, DC, American Psychiatric Association, 1998, pp 1007–1018
35. Basco MR, Bostic JQ, Davies D, Rush AJ, Witte B, Hendrickse W, Barnett V: Methods to improve diagnostic accuracy in a community mental health setting. Am J Psychiatry 2000; 157: 1599–1605
36. Spitzer RL: Psychiatric diagnosis: are clinicians still necessary? Compr Psychiatry 1983; 24:399–411
37. Quality Indicators: Defining and Measuring Quality in Psychiatric Care for Adults and Children: Report of the APA Task Force on Quality Indicators. Washington, DC, American Psychiatric Association, 2002
38. Kerr E, Asch S, Hamilton E, McGlynn EA (eds): Quality of Care for General Medical Conditions: A Review of the Literature and Quality Indicators. Santa Monica, Calif, RAND Health Publications, 2000
39. Rush AJ: DSM-IV mood disorders: final overview, in DSM-IV Sourcebook, vol 4. Edited by Widiger TA, Frances AJ, Pincus HA, Ross R, First MB, Davis W, Kline M. Washington, DC, American Psychiatric Association, 1998, pp 1019–1033
40. Norquist G, Lebowitz B, Hyman S: Expanding the Frontier of Treatment Research. Prevention and Treatment 1999, vol 2
41. Kahn D, Carpenter D, Doherty J, Frances A: The Expert Consensus Guideline Series Treatment of Bipolar Disorder. J Clin Psychiatry 1996; 57(suppl 12A):1–84