Challenges in Operationalizing the DSM-IV Clinical Significance Criterion

Janette Beals, PhD; Douglas K. Novins, MD; Paul Spicer, PhD; Heather D. Orton, MS; Christina M. Mitchell, PhD; Anna E. Barón, PhD; Spero M. Manson, PhD; and the AI-SUPERPFP Team

Background: An explicit clinical significance (CS) criterion was added to many DSM-IV diagnoses in an attempt to more closely approximate the clinical diagnostic process and reduce the proportion of false positives in epidemiological studies. The American Indian Service Utilization, Psychiatric Epidemiology, Risk and Protective Factors Project (AI-SUPERPFP) offered a unique opportunity to examine the success of this effort.

Objective: To determine the impact of distress, impairment, and help-seeking reported in a lay structured interview on concordance with a clinical reappraisal. Further, to test the efficacy of 5 operationalizations of CS on the concordance and prevalence of DSM-IV lifetime disorders.

Design: Completed between 1997 and 2000, a cross-sectional probability sample survey with clinical reappraisal of approximately 10% of participants.

Setting: General community.

Participants: A population-based sample of 3084 members of 2 American Indian tribal groups, who were between the ages of 15 and 54 years and resided on or near their home reservations, were randomly sampled from the tribal rolls and participated in structured psychiatric interviews. Clinical reappraisals were conducted with approximately 10% of the lay-interview participants. The response rate for the lay interview was 75%, and for the clinical reappraisal it was 72%.

Main Outcomes Measures: The AI-SUPERPFP Composite International Diagnostic Interview (CIDI), a culturally adapted version of the CIDI, University of Michigan version. Adapted to assess DSM-IV diagnoses, questions assessing the CS criterion were inserted in all diagnostic modules. The Structured Clinical Interview for DSM-III-R (SCID) was used in the clinical reappraisal.

Results: Most participants who qualified as having AI-SUPERPFP CIDI lifetime disorders reported at least moderate levels of distress or impairment. Evidence of increased concordance between the CIDI and the SCID was lacking when more restrictive operationalizations of CS were used; indeed, the CIDI was very likely to underdiagnose disorders compared with the SCID (false negatives). Concomitantly, the CS operationalizations affected prevalence rates dramatically.

Conclusion: The CS criterion, at least as operationalized to date, demonstrates little effectiveness in increasing the validity of diagnoses using lay-administered structured interviews.
ies and the National Comorbidity Surveys (NCS) suggest that, in any given year, almost 20% to 30% of the population experiences a mental or addictive disorder, while lifetime rates range between 32% and 49%. Clearly, the implications of such estimates for mental health policy in the United States are enormous and have called into question whether these rates accurately reflect the need for treatment.

Critiques of this epidemiological research focus on whether the diagnoses and the criteria defining them successfully differentiate disorder from “problems of living” in other words, whether the thresholds for disorder in such instruments are too low. Hence, in the revisions that occurred between the DSM-III-R and the DSM-IV, an explicit clinical significance (CS) criterion was added to many diagnoses to address at least 1 type of overinclusion: those meeting symptomatic criteria for a disorder but for whom such problems were mild. Although worded somewhat differently across diagnoses, the CS criterion typically asserts that “the symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.”

The conceptual utility and validity of the CS criterion have been much debated; however, empirical examinations of the CS criterion have only recently appeared in the literature. Narrow et al and Regier et al in secondary analyses of the ECA and NCS studies, ascribed CS to participants who had either sought help for specific disorders or reported that these problems interfered with their lives or activities “a lot.” With this operationalization, admittedly limited by the data available in these surveys that predated the DSM-IV, past-year prevalence rates decreased 17% for the ECA (based on the DSM-III-R) and 32% for the NCS (DSM-III-R). Participants meeting both the CS and the symptomatic criteria were more likely than those meeting only symptomatic criteria to have sought services for mental health problems, to have reported either not being able to work or needing to cut down on work, and to have been suicidal. Slade and Andrews, using data from the Australian National Survey of Mental Health and Well-Being (ANSMHWB), found that the inclusion of significant self-reported distress or impairment also decreased the DSM-IV rates of past-month disorder, between 19% for major depressive disorder and 65% for obsessive-compulsive disorder. Controlling for sociodemographics and comorbidity, participants with significant distress/impairment were more likely to report help-seeking, distress, and impairment in other parts of the interview—but only for some diagnoses.

These reanalyses of the ECA, NCS, and ANSMHWB demonstrate that adding CS to the symptomatic criteria substantially lowers prevalence rates and may increase the severity of the resulting diagnoses. However, they also raise additional questions. Each study operationalized CS somewhat differently; also, none was able to compare the impact that multiple operationalizations of CS might have on prevalence. For instance, should only individuals reporting “a lot” of distress or impairment be considered to have met CS or might those reporting moderate or mild distress also qualify as disordered? What role should help-seeking play in this calculus? And what is the degree of overlap among distress, impairment, and help-seeking? Finally, none of these efforts used direct assessments of whether including the CS criterion would increase the concordance between lay- and clinician-administered interviews.

Data from the American Indian Service Utilization, Psychiatric Epidemiology, Risk and Protective Factors Project (AI-SUPERPFP) provided us an opportunity to address these questions. Designed before the World Health Organization’s DSM-IV version of the Composite International Diagnostic Interview (WHO-CIDI 2.1, used by Slade and Andrews) was widely available, the AI-SUPERPFP independently supplemented the CIDI, University of Michigan version (UM-CIDI, used in the NCS) with items necessary to assess DSM-IV criteria in a format conducive to investigating the utility of different operationalizations of CS and the overlap among these constructs. A clinical reappraisal of more than 10% of the sample enabled us to investigate whether the concordance between the lay- and clinician-administered interviews increased with the inclusion of CS.

**METHODS**

**AI-SUPERPFP LAY-INTERVIEW DESIGN AND SAMPLES**

The AI-SUPERPFP sought to estimate the prevalence of psychiatric disorders and health service utilization in 2 American Indian reservation populations. The AI-SUPERPFP methods are described in detail elsewhere; the interview and training manual are available online at http://www.uchsc.edu/aic/ncaainmhr/presentresearch/superpfp.htm. The populations of inference were enrolled members of either 2 closely related Northern Plains tribes or a Southwest tribe who were 15 to 34 years old at the time of development of the sample frame (1997) and who lived on or within 20 miles of their reservations. Once located and found to be eligible, 73.7% (n=1446) from the Southwest tribe and 76.8% (n=1638) from the Northern Plains tribe agreed to participate. Tribal approvals were obtained prior to the project’s beginning. Informed consent was acquired from all participants; with minors, parental/guardian consent was obtained before adolescent assent.

**AI-SUPERPFP CIDI**

As explained in greater detail elsewhere, carefully considered cultural adaptations had already been completed as part of a previous effort conducted between 1992 and 1995. Because that process preceded release of the WHO-CIDI 2.1, this measure was independently augmented and adapted to render it consistent with the DSM-IV.

The present analyses included lifetime diagnoses of the AI-SUPERPFP disorders: panic disorder (PD), generalized anxiety disorder (GAD), posttraumatic stress disorder (PTSD), dysthyemic disorder (DD), major depressive episode (MDE), substance abuse, and substance dependence. Substances included alcohol, sedatives, tranquilizers, stimulants, analgesics, inhalants, marijuana, cocaine, hallucinogens (including peyote), and heroin—each individually assessed and then combined into substance abuse or substance dependence. Additional aggregations included any anxiety disorder (those with GAD, PD, or PTSD), any depressive disorder (MDE or DD), any substance disorder (either abuse or dependence), any anxiety or depressive disorder, and any disorder.
### ASSESSMENT OF CS CONSTRUCTS

**Figure 1** summarizes the operationalizations of CS by Narrow et al. and Regier et al. using the Diagnostic Interview Schedule (DIS)\(^2\) and the UM-CIDI, those of Slade and Andrews\(^2\) using the WHO-CIDI 2.1 in the ANSMHWB, and, finally, the AI-SUPERFPP CIDI measures of distress, impairment, and help-seeking.

Because the ECA-DIS and UM-CIDI predated the DSM-IV, they did not assess CS directly. However, the ECA-DIS used the probe flowchart to distinguish symptoms and syndromes (groupings of symptoms) that were possibly psychiatric from others—a judgment that overlaps but is not synonymous with CS. This set of probes commenced with the question “Did you ever tell a doctor about your [problems just endorsed]?” and continued with questions about other help-seeking, use of medication, and impair-

---

#### Table: DSM-IV Language

| None | Separate Criterion: “The Symptoms Cause Clinically Significant Distress or Impairment in Social, Occupational, or Other Important Areas of Functioning.” | Incorporated in Criterion A: “A Maladaptive Pattern of Substance Use Leading to Clinically Significant Impairment or Distress.” |

#### DIS Items Used in Narrow et al. Reanalyses of Epidemiologic Catchment Area (ECA) Studies

<table>
<thead>
<tr>
<th>Probe Flowchart: Syndrome Level</th>
<th>Diagnosis Not Assessed</th>
<th>Diagnosis Not Assessed</th>
<th>Help-Seeking, Medication Use, or a Lot of Impairment</th>
<th>Help-Seeking, Medication Use, or a Lot of Impairment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Help-Seeking, Medication Use, or a Lot of Impairment</td>
<td>Not Asked</td>
<td>Not Asked</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### UM-CIDI Items Used in Narrow et al. Reanalyses of Baseline National Comorbidity Survey (NCS)

<table>
<thead>
<tr>
<th>Impairment in Important Areas of Functioning: Syndrome Level</th>
<th>How Much Did Your Symptom(s) Interfere With Your Life or Activities—a Lot, Some, a Little, or Not at All?</th>
<th>How Much Did Your Symptom(s) Interfere With Your Life or Activities—a Lot, Some, a Little, or Not at All?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Help-Seeking, Syndrome Level</td>
<td>Did You Ever Tell a, (a) Psychiatrist, (b) Mental Health Specialist, (c) Other Professional, or (d) Take Medications More Than Once Because of X? (4 Questions)</td>
<td>Did You Ever Tell a, (a) Psychiatrist, (b) Mental Health Specialist, (c) Other Professional, or (d) Take Medications More Than Once Because of X? (4 Questions)</td>
</tr>
</tbody>
</table>

#### WHO-CIDI 2.1 Items Used in Slade and Andrews\(^2\) Reanalyses of Australian National Survey of Mental Health and Well-Being (ANSMHWB)

<table>
<thead>
<tr>
<th>Distress</th>
<th>Have You Ever Been Very Upset With Yourself for Feeling Worried, Tense, or Anxious for Long Periods of Time?</th>
<th>Have You Ever Been Very Upset With Yourself for Having Problems That Occurred as a Result of [Event]?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impairment in Important Areas of Functioning</td>
<td>Did the Problems That Occurred as a Result of [Event] Ever Interfere With Your Life or Activities a Lot?</td>
<td>Did Any Period Lasting 2 Weeks or Longer Seriously Interfere With Your Ability to Do Your Job, Take Care of Your House or Family, or Take Care of Yourself?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Probe Flowchart: Syndrome Level</th>
<th>Help-Seeking, Medication Use, or a Lot of Impairment</th>
<th>Help-Seeking, Medication Use, or a Lot of Impairment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Help-Seeking, Medication Use, or a Lot of Impairment</td>
<td>Not Asked</td>
<td>Not Asked</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Probe Flowchart: Symptom Level</th>
<th>Not Asked</th>
<th>Not Asked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Help-Seeking, Medication Use, or a Lot of Impairment</td>
<td>Asked for 2 Screeners</td>
<td>Asked for 1 Screener</td>
</tr>
</tbody>
</table>

#### American Indian Service Utilization, Psychiatric Epidemiology, Risk and Protective Factors Project (AI-SUPERFPP) CIDI Items

<table>
<thead>
<tr>
<th>Distress</th>
<th>How Much Did X Ever Bother or Upset You? Would You Say Not At All, Some, or a Lot?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impairment in Important Areas of Functioning</td>
<td>How Much Did X Ever Interfere With Your Life or Activities? Would You Say Not At All, Some, or a Lot?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Help-Seeking</th>
<th>Did You Ever Talk to a (a) Mental Health Specialist or (b) Medical Person? (c) Did You Ever Receive Treatment for X? (d) Did You Ever Take Medicine That Was Prescribed by a Doctor for X? (e) Did You Ever Take Any Other Herbs, Roots, or Remedies for Problems Like These?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Did You Ever Talk to a Healer or Spiritual or Traditional Leader About X? (b) Did You Ever Take Medicine That Was Given to You by a Medicine Man or Woman for X? (c) Did You Ever Take Any Other Herbs, Roots, or Remedies for Problems Like These?</td>
<td></td>
</tr>
</tbody>
</table>

---

Figure 1. Definitions of clinical significance criteria in DSM-IV and items used in 3 studies to operationalize clinical significance construct. Shaded cells indicate the additions to the diagnostic algorithms in Narrow et al\(^3\) and Slade and Andrews.\(^4\) DIS indicates Diagnostic Interview Schedule; UM-CIDI, Composite International Diagnostic Interview, University of Michigan version; and WHO, World Health Organization.
ment in terms of the specific symptom/syndrome. When the probe flowchart was assessed at the symptom level, only individuals with probable psychiatric symptoms advanced in the diagnostic modules. As shown in Figure 1, the ECA-DIS used the probe flowchart at the symptom level for PD and DD, and thus the probe flowchart was a necessary component of these diagnoses. On the other hand, the probes for MDE and substance disorders were asked at the syndrome level and have been only recently introduced into the prevalence calculus in the work by Narrow et al and Regier et al. Similarly, the UM-CIDI assessed help-seeking, medication use, and impairment at the syndrome level; again, these data did not influence prevalence estimates until the recent secondary analyses. In the WHO-CIDI 2.1 algorithms, an instrument designed to assess DSM-IV, the probe flowchart was used at the symptom level only for MDE and DD; otherwise, it was used at the syndrome level. The distress and impairment questions specific to the DSM-IV were asked for GAD, PTSD, and MDE (impairment only).

In the AI-SUPERPFP CIDI, identical items assessed self-reported distress, impairment, and help-seeking at the syndrome level across diagnoses. Impairment and help-seeking questions were patterned after those of the UM-CIDI but altered slightly to maximize cultural validity. For instance, whereas the UM-CIDI included “a little” as a response option in its impairment items, this was dropped in the AI-SUPERPFP because focus groups suggested that participants were unlikely to reliably differentiate between “some” and “a little.” Help-seeking questions reflected the service ecology of reservation residents by including specific types of service providers (eg, community health representatives as medical personnel) and directly assessing use of traditional healing resources.

OPERATIONALIZATIONS OF CS

Using the AI-SUPERPFP data, 5 operationalizations of CS were compared. The first (CS0) excluded the CS criterion. The next 3 operationalizations built upon one another, ranging from less (CS1) to more (CS3) restrictive, and focused on degrees of self-reported distress and impairment, adhering closely to the DSM-IV CS language. The final operationalization (CS4) included both help-seeking and impairment and most closely mimicked the probe flowchart.

Operationalization CS0: no assessment of clinically significant distress or impairment. Diagnoses were based on symptomatic criteria only.

Operationalization CS1: “a lot or some” impairment or distress. A disorder was considered clinically significant if the participant reported “a lot” or “some” distress or impairment. This follows Wakefield and Spitzer’s suggestion that those expressing either moderate or severe distress or impairment should be considered “true” positives.

Operationalization CS2: “a lot” of distress or impairment or “some” of both. As a variation of the definition of moderate disability, we included an operationalization whereby experiencing “a lot” of distress or impairment or some deleterious effects in multiple domains merited a diagnosis. This operationalization represented a middle ground between CS1 and CS3.

Operationalization CS3: “a lot” of distress or impairment. Here participants reporting “a lot” of distress or impairment were considered to meet the definition of CS. This operationalization was closest to that of Slade and Andrews.

Operationalization CS4: help-seeking or “a lot” of impairment. This operationalization most closely matches the work of Narrow et al and Regier et al, albeit at the syndrome level. Seeing or talking to a mental health provider or other medical personnel about the psychiatric symptoms or having been hospitalized for these problems constituted help-seeking.

CLINICAL REAPPRAISAL

The AI-SUPERPFP included a clinical reappraisal of approximately 10% (n=335) of participants, who were re-interviewed by psychiatrists or clinical psychologists. This component was designed to assess the concordance between the AI-SUPERPFP CIDI and the Structured Clinical Interview for DSM-III-R, nonpatient version (SCID). Approximately 75% of the clinical reappraisal sample was randomly chosen based on a positive CIDI diagnosis of the 3 most common disorders: MDE, PTSD, or alcohol abuse/dependence. The remaining sample, also randomly selected, did not qualify for any AI-SUPERPFP diagnosis. Because the greatest source of error between lay and clinical interviews is commonly found in those who have some but not all of the symptoms required for a diagnosis (subthreshold cases), approximately half of those in the no-disorder group endorsed significant levels of depressed, anxious, or irritable symptoms on a checklist independent of the CIDI, with the remainder having few or none of these symptoms. The SCID was adapted to allow for changes between the DSM-III-R and the DSM-IV, including the CS criterion. The 8 clinician interviewers had extensive clinical experience (more than 15 years on average) and had worked in American Indian communities. Before entering the field, each demonstrated a high level of interrater reliability (κ=.80) in a series of videotapes coded by an expert panel, and they also performed supervised interviews with members of the local American Indian community. Furthermore, all clinical reappraisals were audi-taped and reviewed by master clinicians for quality assurance purposes. The response rate for the clinical reappraisal sub-study was 72.3% and was similar to that for the CIDI-disordered and -nondisordered participants. An average of 120 days elapsed between lay and clinical interviews. There was no association between the time elapsed between interviews and agreement between the CIDI and the SCID. Clinicians were blind to participants’ CIDI diagnostic status.

STATISTICAL METHODS

Variable construction and noninferential analyses were completed using SAS, SPSS, and Stata. First, to better understand the patterns of distress, impairment, and help-seeking among participants meeting symptom criteria for AI-SUPERPFP CIDI disorders, frequencies and associated confidence intervals for these constructs are presented in Table 1 for each disorder. Two Venn diagrams (Figure 2) depict the overlap among these constructs for those qualifying for at least 1 disorder. As a second step, concordance between the AI-SUPERPFP CIDI and the SCID was evaluated by assessing the numbers of true and false positives and negatives generated by the CIDI when the SCID was considered the gold standard, as well as following the standard of set statistics: (1) Cohen κ; (2) sensitivity and specificity; (3) positive and negative predictive values calculated using the Bayes rule; and (4) the McNemar χ² test (a measure of bias). Although not an assessment of concordance, the Global Assessment of Functioning scores, determined by the clinicians during the reappraisal, provided a measure of severity of disorder and are included in Table 2. Finally, Table 3 presents the differential prevalences, with associated 95% confidence intervals, of specific and aggregated disorders across the 5 CS operationalizations using the AI-SUPERPFP CIDI. The results in Table 1 and Table 3 provide inferences to the populations and were conducted in Stata using sample and nonresponse weights. Since the concordance analyses in Table 2 focus on relative functioning of instruments, unweighted estimates were deemed acceptable and offered the added benefit of being able to provide the actual numbers of participants in various cells in the table.
RESULTS

CS CONSTRUCTS

Table 1 presents the self-reported distress, impairment, and help-seeking for participants meeting the symptom criteria for DSM-IV lifetime disorders in AI-SUPERPFP. When considering distress, among those with any AI-SUPERPFP CIDI disorder, 46.5% reported being upset or bothered “a lot” (distressed) by their symptoms. Between 35% and 50% of those with depressive or anxiety disorders reported such levels of distress. Among those...
with substance problems, a majority with dependence reported high levels of distress compared with about one quarter with substance abuse. When reports of “some” distress were included, more than 90% of participants who were assigned a diagnosis other than substance abuse reported being upset or bothered by their symptoms. Overall, an additional 39.9% (95% confidence interval, 36.4%-46.5%) with any disorder reported being bothered or upset “some” when compared with “a lot,” with the difference somewhat larger for depressive disorders (56.2%) than either anxiety disorders (44.4%) or substance disorders (40.2%).

Similarly, for impairment, one third of participants with any disorder reported that their symptoms interfered with their lives and activities “a lot.” Those with substance dependence ranked highest, while those with substance abuse ranked lowest. The difference between “a lot” and “some” was 49.2% overall and followed patterns similar to those seen for distress.

Turning to help-seeking, almost half of the participants qualifying for 1 of the AI-SUPERPFP disorders had sought biomedical help for their symptoms; about one third had sought help from traditional healing sources. Generally, the pattern of seeking help from traditional sources mirrored that from biomedical sources. Combining traditional with biomedical sources of help-seeking increased the rates by 9.6%.

### CONCORDANCE ANALYSES

Overall, 78% of the clinical reappraisal sample was judged by the SCID interviewers to have a disorder; without considering CS (that is, CS0), the CIDI designated 70% of this select sample as having at least 1 DSM-IV disorder.

### Table 2. Concordance Between the AI-SUPERPFP CIDI and the SCID for Lifetime Disorders*

<table>
<thead>
<tr>
<th>Constructs That Have Been Used to Define Clinical Significance</th>
<th>True Positives</th>
<th>False Negatives</th>
<th>False Positives</th>
<th>True Negatives</th>
<th>Cohen’s k</th>
<th>Sensitivity, %</th>
<th>Specificity, %</th>
<th>PPV, %</th>
<th>NPV, %</th>
<th>Bias, %</th>
<th>GAF, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distress: A lot</td>
<td>128</td>
<td>127</td>
<td>8</td>
<td>65</td>
<td>0.25</td>
<td>50</td>
<td>89</td>
<td>93</td>
<td>37</td>
<td>117.6</td>
<td>64.8</td>
</tr>
<tr>
<td>Distress: Some or a lot</td>
<td>193</td>
<td>62</td>
<td>15</td>
<td>58</td>
<td>0.45</td>
<td>76</td>
<td>79</td>
<td>92</td>
<td>52</td>
<td>49.1</td>
<td>65.2</td>
</tr>
<tr>
<td>Impairment: A lot</td>
<td>90</td>
<td>165</td>
<td>8</td>
<td>65</td>
<td>0.14</td>
<td>35</td>
<td>89</td>
<td>91</td>
<td>31</td>
<td>155.5</td>
<td>62.8</td>
</tr>
<tr>
<td>Impairment: Some or a lot</td>
<td>182</td>
<td>73</td>
<td>15</td>
<td>58</td>
<td>0.40</td>
<td>71</td>
<td>79</td>
<td>91</td>
<td>48</td>
<td>59.5</td>
<td>64.2</td>
</tr>
<tr>
<td>Help-seeking: Biomedical</td>
<td>117</td>
<td>138</td>
<td>7</td>
<td>66</td>
<td>0.22</td>
<td>46</td>
<td>90</td>
<td>93</td>
<td>36</td>
<td>129.5</td>
<td>62.4</td>
</tr>
<tr>
<td>Help-seeking: Traditional</td>
<td>68</td>
<td>187</td>
<td>5</td>
<td>68</td>
<td>0.10</td>
<td>27</td>
<td>93</td>
<td>92</td>
<td>30</td>
<td>180.2</td>
<td>64.5</td>
</tr>
<tr>
<td>Help-seeking: Either biomedical or traditional</td>
<td>140</td>
<td>115</td>
<td>7</td>
<td>66</td>
<td>0.30</td>
<td>55</td>
<td>90</td>
<td>94</td>
<td>40</td>
<td>106.6</td>
<td>63.5</td>
</tr>
</tbody>
</table>

Abbreviations: AI-SUPERPFP, American Indian Service Utilization, Psychiatric Epidemiology, Risk and Protective Factors Project; CIDI, Composite International Diagnostic Interview; CS, clinical significance; GAF, Global Assessment of Functioning; NPV, negative predictive value; PPV, positive predictive value; SCID, Structured Clinical Interview for DSM-III-R; WHO-CIDI 2.1, World Health Organization’s DSM-IV version of the CIDI.

*Statistical tests between operationalizations were inappropriate because they do not represent independent samples. Similar conclusions were reached whether or not the data were weighted back to the lay sample or if the data were restricted to those disorders for which DSM-IV includes specific CS language; for ease of presentation, the unweighted results of all disorders are reported.

†Calculated using the Bayes rule (that is, using the SCID prevalence weighted back to the full population).
‡Calculated using the McNemar $\chi^2$ test.
§Mean score for those with a disorder under this operationalization.
¶The AI-SUPERPFP approximation of WHO-CIDI 2.1 operationalization.
||The AI-SUPERPFP approximation of Narrow and Regier operationalization.
The concordance between the AI-SUPERPFP CIDI and clinical reappraisals with the SCID is presented in Table 2. Similar to others’ reports,90-92 agreement between these clinical and lay methods of case ascertainment was modest. However, the hypothesis tested here was whether inclusion of the CS operationalizations in the lay-interview data would more closely approximate clinical diagnoses.

Focusing first on distress, impairment, and help-seeking constructs, the $k$ values were lower for the more restrictive response patterns. Sensitivity assesses the ability of the CIDI to record a positive diagnosis for SCID-defined cases; here, a 26% decrease in sensitivities arose when we designated those having “a lot” of distress as meeting CS compared with including “some or a lot.” In contrast, the specificity or ability of the CIDI to identify the SCID-defined noncases increased 10% between the same definitions. The positive predictive values indicated that the vast majority of the CIDI-defined cases also received SCID diagnoses. The negative predictive values revealed that as the CS criteria became more restrictive, a greater percentage of CIDI-defined noncases actually received SCID diagnoses. The bias was more severe as the requirements increased for being labeled distressed, impaired, and a help-seeker. Finally, although the Global Assessment of Functioning scores decreased with the more restrictive operationalizations, the differences between them were minimal.

To this point, we have discussed distress, impairment, and help-seeking as separate constructs. Figure 2 depicts the overlap among distress, impairment, and help-seeking for participants qualifying for any of the AI-SUPERPFP disorders. When reports of “a lot” of either distress or impairment were required, almost one third (32.3%) of participants meeting symptomatic criteria failed to report sufficient distress/impairment or help-seeking. Less than 20% reported all 3, and 16.4% reported help-seeking only. Considering the overlap when participants who reported “some” or “a lot” of distress or impairment were included, many fewer reported no distress, impairment, or help-seeking (8.1%). Here, the largest categories included participants who reported all 3 indicators of CS (42.4%) and those who reported both distress and impairment (35.7%). Thus, the extent of this overlap varied dramatically by the inclusiveness of the responses considered as markers of distress and impairment.

Table 2 also examines the concordance statistics for 4 methods of combining these distress, impairment, and help-seeking measures with varying levels of response inclusiveness, informed by previous work in this area. Thus, CS3 (“a lot” of distress or impairment) approximated the definition found in Slade and Andrews,23 while CS4 (help-seeking or “a lot” of impairment) was closest to the approaches of Narrow et al4 and Regier et al.21 Operation-
alizations CS1 (“some” distress or impairment) and CS2 (“a lot” of distress or impairment or “some” of both) followed Wakefield and Spitzer’s suggestions and included moderate levels of distress and impairment. Once again, as the operationalizations became more restrictive, the $\kappa$ values decreased, the sensitivities decreased more dramatically than the specificities increased, the positive predictive values remained stable while the negative predictive values decreased dramatically, the biases increased, and the Global Assessment of Functioning scores remained quite stable. These results indicate that CS was not the source of disagreement between the AI-SUPERPFP CIDI and the SCID.

**PREVALENCE ACROSS DIFFERENT OPERATIONALIZATIONS OF CLINICALLY SIGNIFICANT DISTRESS OR IMPAIRMENT**

Next we turn to the effect the various 5 operationalizations of CS had on prevalence estimates, focusing first on “any disorder.” Operationalization CS3 is the most conservative: the prevalence rate was 50.1% of that when no CS criterion was used (22.8% using CS3 compared with 45.5% with no CS criterion). Similarly, the relative rate for CS4 was 98.0%, 77.4% for CS2, and 88.3% for CS1. When the “any disorder” category was restricted to only diagnoses with an explicit CS criterion in the DSM-IV (GAD, PTSD, MDE, and DD), the relative rates were less dramatic and ranged between 71.4% for CS3 and 95.6% for CS1. Figure 3 demonstrates the relative diminution by disorder across the various operationalizations.

Imposing the CS criterion had the greatest impact for substance abuse. Although the DSM-IV implies that CS is less relevant to defining substance dependence and PD, the patterns in these instances were quite similar to those of other disorders. When compared with the secondary analyses of the NCS of Narrow et al and Regier et al (see Table 2 of Narrow et al), the AI-SUPERPFP prevalence rates applying CS4 were reduced to a greater extent than was observed in the NCS (PD reduced 35.3% [100%-2.2%/3.4% = 35.3%] compared with 22.7%; GAD, 34.5% compared with 17.6%; MDE, 40.0% compared with 36.6%; DD, 45.5% compared with 28.0%; and substance disorders, 50.6% compared with 33.9% for the AI-SUPERPFP and the NCS, respectively).

**COMMENT**

The concern that psychiatric epidemiological methods, especially those relying on structured protocols administered by lay interviewers, should reflect the clinical expertise inherent in the diagnostic process of the DSM was a driving force behind the deliberate inclusion of the CS criterion in the DSM-IV. Subsequently, 2 basic questions have informed epidemiological work on CS. First, were the unexpectedly high rates reported by the ECA and the NCS inflated by “false positives,” that is, by participants meeting the symptomatic criterion for disorder but for whom such problems were relatively inconsequential? Second, given that a central goal of studies like the ECA and the NCS was to project treatment need,
did adding CS to the diagnostic algorithm render the estimates more appropriate for mental health policy planning? The data presented here, and our experiences in cross-cultural settings generally, provide additional insight into these questions and further inform the debate about CS in preparation for the DSM-V.

The AI-SUPERPFP DSM-IV rates of any lifetime disorder were 45.5% when CS was not considered. When the DSM-III-R rates were compared directly with those of the NCS, the prevalence of disorders was somewhat higher (a range of 2% to 10% based on tribe and sex) than those in the NCS. The next question, then, was whether false positives inflated these rates. Using these methods within this cultural context, false negatives rather than false positives were the major source of discrepancy between lay- and clinician-administered interviews. This finding does not appear to be specific to the AI-SUPERPFP, having also been reported with the ECA data. Within the baseline NCS, MDE exhibited significant levels of false positives when compared with the SCID; the bias for most disorders was in the opposite direction, although statistically significant only for simple phobias. The recent NCS-Replication appears to have addressed the issue of false positives for lifetime but not current MDE when compared with clinical reappraisals. Thus, with the exception of the baseline NCS MDE rates, lay interviews appear to underestimate rather than overestimate lifetime disorder when compared with clinical interviews—the reverse of the bias the CS criterion was designed to address. Indeed, Spitzer and Wakefield anticipated that the CS criterion might dramatically increase the false negatives while having little influence on the false positives. These findings support their hypothesis.

Thus, if the SCID were considered a true gold standard, the conclusion that the CS criterion should be ignored would be reasonable, for the AI-SUPERPFP at least. However, we are not yet asserting the primacy of one method of case ascertainment over the other. Rather, we consider both lay- and clinician-administered interviews to be potentially biased. For instance, underreporting may be more common in the lay interviews, while clinicians may be more prone to attribute disorder to “normal” behaviors. Furthermore, biases inherent in the measurement of DSM-defined diagnoses in these cultural contexts may differentially affect lay- and clinician-administered interviews. While in-depth investigations of the relative validity of the lay- and clinician-administered protocols will be a focus of our investigative energies in coming years, the current analyses have implications for others using DSM-IV definitions of disorder.

As seen in Figure 1, the operationalizations of CS to date have differed considerably. Neither the ECA-DIS nor the UM-CIDI was designed to assess DSM-IV disorders; thus, their definitions preceded the explicit inclusion of the CS criterion in the DSM-IV. At the same time, their probe-flowchart data and the resulting definitions of “probable” psychiatric symptoms drove much of the pre-DSM-IV debate about CS. The WHO-CIDI 2.1 was designed for the DSM-IV and, in many senses, represents a compromise between the ECA-DIS and the UM-CIDI approaches with its use of the probe flowchart, mostly at the syndrome level, and with individual items also assessing clinical significance for GAD, PTSD, and MDE. Figure 1 illustrates considerable variation in the measurement of CS; further, the items often do not closely match the DSM-IV language. Survey methodologists have consistently demonstrated the large impact of even small differences in wording. As a limitation to the joint ECA/NCS analyses, Narrow et al pointed out that the ECA impairment question included “a lot” in the stem whereas NCS did not. Thus, it is unknown whether participants answering “some” to the NCS question might choose “yes” or “no” in the ECA version. As shown in Table 1 and Table 3, the AI-SUPERPFP data suggest that such differences may be substantial.

Figure 1 also highlights the differential inclusion of help-seeking in the diagnostic calculus across instruments. Table 1 and Figure 2 provide data on the prevalence of help-seeking in the AI-SUPERPFP samples and the overlap among help-seeking, distress, and impairment. Individuals seeking help for their symptoms are often distressed or impaired; however, the inclusion of help-seeking may carry with it an assumption that adequate services are available and known to be efficacious and acceptable to community members. In preparation for the AI-SUPERPFP, both the ECA-DIS and the UM-CIDI were submitted to focus group review, and concerns were raised about the use of the ECA-DIS probe flowchart with its help-seeking stem question. In particular, our informants suggested that many American Indians with emotional problems have learned from hard experience that local service providers are few in number and often lack the expertise or training to treat such matters. Also, many American Indians who suffer from mental disorders seek treatment from traditional healing sources. Even at this earlier stage in our research, therefore, serious concerns were raised about the use of a help-seeking question as a conditional definition of probable symptoms.

Before concluding, limitations of the current work deserve mention. The samples from which these data were derived limit the inferences drawn. These data were restricted to American Indian participants and, even then, represented only 3 of more than 300 federally recognized American Indian tribes; participants were restricted to members living on or near their reservations and covered a limited age range. Further, the analyses were limited; in particular, the concordance analyses required an assumption that the SCID be considered a gold standard, and thus the estimations of the sensitivities and specificities were likely biased to some degree. Finally, we did not assess the viability of other constructs such as “harmful dysfunction” to explain the remaining false positives; an investigation of the cultural definitions of such constructs in American Indian communities is strongly recommended.

Even with these limitations, the current work informs ongoing debates about the CS criterion and other definitions of probable psychiatric symptoms. As others have noted, the lack of consistency with which the CS criterion is applied in the DSM-IV is unsettling and may
reflect the ambivalence of the field about this construct. In the absence of biological markers, most diagnoses must superimpose a threshold on dimensions of psychopathology. Further inclusion of thresholds based on disability threatens to make the diagnostic calculus unmanageable and, on the basis of the data reported here, may have limited value. As previously argued, and as operationalized now in the International Classification of Diseases, 10th Revision, we suggest the authors of the DSM seriously consider uncoupling assessments of disability from diagnosis, which would serve to emphasize, in a slightly different manner than do CS criteria, that diagnosis should not in itself be equated with medical necessity.

Submitted for Publication: October 20, 2003; final revision received December 30, 2003; accepted April 21, 2004.

Correspondence: Janette Beals, PhD, American Indian and Alaska Native Programs, University of Colorado Health Sciences Center, MS F800, PO Box 6508, Aurora, CO 80045-0508 (jan.beals@uchsc.edu).

Additional Authors/The AI-SUPERPFP Team: Cecelia K. Big Crow, Dedra Buchwald, MD, Buck Chambers, Michelle L. Christensen, PhD, Denise A. Dillard, PhD, Karen DuBray, Paula A. Espinoza, PhD, Candace M. Fleming, PhD, Ann Wilson Frederick, Diana Gurley, PhD, Lori L. Jervis, PhD, Shirlene M. Jim, Carol E. Kaufman, PhD, Ellen M. Keane, Suzell A. Klein, Denise Lee, Monica C. McNulty, Denise L. Middlebrook, PhD, Laurie A. Moore, Tilda D. Nez, Ilena M. Norton, MD, Carlette J. Randall, Angela Sam, James H. Shore, MD, Sylvia G. Simpson, MD, and Lorette L. Yazzie.

Funding/Support: This study was supported by the following grants from the National Institutes of Health (NIH), Bethesda, Md: R01 MH48174 (Dr Manson) and R01 AA14817 (Dr Beals) and R01 AA13420 (Dr Beals).

Acknowledgment: The AI-SUPERPFP would not have been possible without the significant contributions of many people. The following interviewers and computer/data management and administrative staff supplied energy and enthusiasm for an often difficult job: Amelia T. Begay, Cathy A. E. Bell, Mary Cook, Helen J. Curley, Mary C. Davenport, Rhonda Wiegm Anick, Marvine D. Douville, Geneva Emhoohal, Fay Flame, Roslyn Green, Billie K. Greene, Jack Herman, Tamara Holmes, Shelly Hubing, Cameron R. Joe, Louise F. Joe, Chery 1. Martin, Jeff Miller, Robert H. Moran, Jr, Natalie K. Murphy, Ralph L. Roanhorse, Margo Schwab, PhD, Jennifer Settlemire, Donna M. Shangreaux, Matilda J. Shorty, Selena S. S. Simons, Jennifer Truel, Lori Trullinger, Jennifer M. Warren, Theresa (Dawn) Wright, Jenny J. Yazzie, and Sheila A. Young. We would also like to acknowledge the contributions of the Methods Advisory Group: Margarita Alegría, PhD, Evelyn J. Bromet, PhD, Dedra Buchwald, MD, Steven G. Heeringa, PhD, Ronald Kessler, PhD, Peter Guarnaccia, PhD, R. Jay Turner, PhD, and William A. Vega, PhD. William E. Narrow, MD, Tim Slade, PhD, and Gavin Andrews, MD, are gratefully acknowledged for excellent suggestions based on a review of the manuscript before submission. We are also indebted to the ARCHIVES reviewers, whose comments greatly improved the manuscript. Finally, we thank the tribal members who so generously answered all the questions asked of them.

REFERENCES


40. Brugha TS, Jenkins R, Taub N, Meltzer H, Bebbington PE. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). Psychol Med. 2001;31:1001-1013.


42. Robins LN. Epidemiology: reflection on testing the validity of psychiatric interviews. Arch Gen Psychiatry. 1985;42:918-924.


