A Comparison of Diagnostic Interviews for Depression in the Stirling County Study

Challenges for Psychiatric Epidemiology

Jane M. Murphy, PhD; Richard R. Monson, MD, DSc; Nan M. Laird, PhD; Arthur M. Sobol, MA; Alexander H. Leighton, MD

Background: High prevalence rates in psychiatric epidemiologic studies raise questions about whether data-gathering procedures identify transient responses rather than clinical disorders. This issue is explored relevant to depression using data from the Stirling County Study.

Methods: The study’s customary method, the DPAX (DP for depression and AX for anxiety) was compared with the Diagnostic Interview Schedule (DIS), both of which were administered to a sample of 1396 subjects selected in 1992. Reasons for discordance were analyzed, and demographic correlates of responses to questions about dysphoria were examined. These lay-administered interviews were then compared with clinician-administered interviews that used the Structured Clinical Interview for DSM-III-R (SCID) with 139 subjects. The $\kappa$ statistic and logistic regression were used for statistical assessment.

Results: For the level of agreement between the DPAX and the DIS for current and lifetime depression, $\kappa = 0.40$ and $\kappa = 0.33$, respectively. Subjects diagnosed only by the DPAX tended to have less education than those diagnosed only by the DIS. Some idioms for dysphoria seemed to work better than others. Using SCID interviews as a clinical standard, the DPAX had 15% sensitivity and 96% specificity and the DIS had 25% sensitivity and 98% specificity.

Conclusions: Comprehension of an interview can be improved by using multiple questions for dysphoria and a simpler mode of inquiry. Clinician-administered interviews tend to corroborate disorders identified in lay-administered interviews but suggest that survey methods underestimate prevalence. Further research is needed to evaluate the validity of both types of interviews, but evidence from a 16-year follow-up evaluation indicates that depression diagnosed by the DPAX is a serious disorder in terms of morbidity and mortality.

Arch Gen Psychiatry. 2000;57:230-236

In comparing results from the Epidemiologic Catchment Area (ECA) program and the National Comorbidity Survey, Regier et al concluded that psychiatric epidemiology “faces new challenges.” Because high rates were reported for the National Comorbidity Survey, it was suggested that survey methods may mistake “normal and transient responses” for “clinical disorders.” We address some of the issues involved in this challenge using information from the Stirling County Study.

See also pages 209, 217, 223, and 227

In recent data gathering, we included the depression module of the Diagnostic Interview Schedule (DIS) of the ECA, along with the depression component of version 2 of our customary method, the DPAX-2 (DP for depression and AX for anxiety). Despite differences in the questions asked and the style of inquiry, prevalence rates for current depression were generally similar. Recognizing that similarity in rates does not necessarily mean that the same individuals are counted, we analyzed the agreement for both current and lifetime depression.

To compile information about the content of questions and ways of asking them that might lead to improving the comprehensibility and reliability of lay-administered interviews, we studied the reasons for disagreement between the DPAX-2 and the DIS. We also investigated responses to key questions to see if they seemed to be understood equally well by different segments of the population.

To evaluate whether disorders identified in community surveys are similar to those recognized by clinicians, we report results based on the Structured Clinical Interview for DSM-III-R (SCID). Because the clinical significance of a disorder is of-
PARTICIPANTS AND METHODS

DESIGN AND SAMPLING

The Stirling County Study is a long-term investigation of psychiatric epidemiology in a general population located in Atlantic Canada. The design of the study is a combination of repeated cross-sectional sample surveys (1952, 1970, and 1992) and cohort follow-up investigations.17

This report deals mainly with data gathered in the 1990s, when both the DPAX-2 and the DIS were administered to the same subjects. The focus is on the 1396 subjects (86% completion rate) selected as a representative sample of adults in 1992.5,11,18 In terms of demographic characteristics, 54% of the sample members (n = 738) were women, 43% (n = 604) were younger than 45 years, and 53% (n = 746) had received less than an 11th grade education.

The DPAX-2 and DIS interviews were carried out in the homes of subjects by graduate students who received 10 days of training and practice. Instruction was provided by the staff of the Stirling County Study, one of whom had been trained at the DIS headquarters at Washington University, St Louis, Mo. Subjects gave informed consent for these interviews as well as for those using the SCID.

The goal in selecting subjects for SCID interviews was to create a matrix balanced for age and sex in which there would be a concentration of subjects who were positive for a condition according to the DPAX-2 and/or DIS and in which there would be representation of noncases as well.10 Diagnoses used in the selection included depression, anxiety disorders, and alcohol abuse. The DPAX-2 and DIS records were systematically searched until an appropriate person was found for a given category of age, sex, and diagnosis. By this means, 98 sample subjects and 43 follow-up subjects were chosen and interviewed (92% completion rate).

The SCID interviews were carried out by psychiatrists and clinical psychologists who were blind to the survey results but who had been trained in the administration of the nonpatient version.17 These interviews were conducted in the homes of the subjects on average 2 years after the DPAX-2 and DIS interviews. Five subjects were diagnosed by the clinicians as having had a SCID depression that occurred only after the lay interviewing had been concluded. For purposes of comparability between the clinician and lay interviews, these subjects were counted in the analysis as never having had a depression. Except for this adjustment, the SCID analysis deals with lifetime diagnoses of major depressive episode (MDE) and dysthymia.

DPAX AND DIS INTERVIEW METHODS

The DPAX-1 method was designed for our 1952 study as a means of estimating “need for psychiatric attention.”6,5 The DPAX-2 was designed to take account of additions that were made to the interview schedule for the 1970 study and thereafter to overcome the psychometric weakness of having only one question about dysphoria and only one about functional impairment.10 The DPAX-2 is used in this article except for a review of follow-up findings from the 1952 study. The DIS module for depression was administered after the portion of the interview dealing with the DPAX-2. The DIS was designed to implement the definitions given in DSM-III.20

Both the DPAX-2 and DIS follow a diagnostic algorithm that establishes the presence of the “essential features” of the depressive syndrome, addresses the completeness of the syndrome through requiring a certain number of “associated symptoms,” and determines that criteria for a minimum duration have been met. However, there are important differences between the 2 methods in the questions asked for these diagnostic components, especially in the number and range of options offered.

RESULTS

BACKGROUND FINDINGS AND COMPARISON OF THE DPAX WITH THE DIS

We recently reported that the current prevalence of depression in Stirling County remained stable at slightly more than 5% across the samples of 1952, 1970, and 1992.11 In the first 2 samples, men and women had quite similar rates, and for both, prevalence increased with age. By 1992, a redistribution had occurred, which indicated higher prevalence among younger women. The DIS prevalence of current depression in 1992 was similarly close to 5%, and the DIS rates by sex and age showed considerable comparability with the DPAX-2, especially among women (Table 1). Lifetime rates were not as similar, since the DIS identified a larger proportion of past episodes than the DPAX-2. Despite similarity in current rates, case-by-case analysis indicated that the level of agreement as measured by the κ statistic was only fair (Figure 1). Where past depression was concerned, agreement was poor.

REASONS FOR DISAGREEMENT

For each method, each component of the diagnostic algorithm used by the other method was associated with one or more disagreements regarding both current and past depression (Table 2). A larger proportion of disagreements regarding essential features pertained to the DIS, with its more complex approach, than to the DPAX-2, with its 3 optional questions. On the other hand, more disagreements regarding associated symptoms involved the DPAX-2, with its stringent requirement for each of 3 specific types of disturbance, in contrast to the larger number of options available in the DIS.

While the analysis presented does not distinguish between MDE and dysthymia, 24 (65%) of the 37 current MDE cases were also identified by the DPAX-2. Con-
The DIS uses a complex opening question for MDE that combines the essential features of dysphoria and anhedonia with minimum duration: “In your lifetime, have you ever had 2 weeks or more during which you felt sad, blue, depressed, or when you lost all interest and pleasure in things that you usually cared about or enjoyed?” This is followed by a similar but simpler question for dysthymia: “Have you had 2 years or more in your life when you felt depressed or sad most days, even if you felt OK sometimes?” While the DIS offers another chance for giving evidence about essential features, it was rarely used. Thus, on the whole, the DIS requires a positive response to one question about essential features for each diagnosis.

The DPAX-2, on the other hand, uses 3 separate questions for dysphoria as the essential feature: “Have there been times when you felt low and hopeless?” “Do you sometimes wonder if anything is worthwhile any more?” “Do you feel in good spirits?” A negative response to the last question is interpreted as positive evidence of “poor spirits.” Three optional ways of describing dysphoria are offered, with a positive response required for one.

The DIS inquires about the 8 types of associated symptoms given in DSM-III and requires 4. For associated symptoms, the DPAX-2 asks about disturbances of appetite, sleep, and energy, and requires that each be present. For “duration,” the DIS uses predetermined durations embedded in the questions about essential symptoms and later gathers information about time of occurrence. The DPAX-2 offers an opportunity for describing duration as it pertains to individual cases by using open-ended questions about when the essential symptom became noticeable, whether it was still bothersome at the time of interview, and, if not, when it ceased to be troubling. A minimum duration of 1 month is applied later by computer.

For the DPAX-2, the term current depression refers to the episodes present at the time of interview. For the DIS, current depression is defined as MDE present at the time of interview or within the preceding month, together with dysthymia. The reason for combining the episodic and chronic types of DIS depression is to provide a parallel to the DPAX-2, which identifies episodes with varying durations and different levels of severity.

The DIS has additional criteria for the psychiatric relevance of the symptoms, temporal clustering of features and symptoms as in a spell, duration and severity of the spell, and whether the spell occurred only as part of bereavement. In keeping with DSM-III, DSM-III-R, and DSM-IV, the underlying concept in the DIS questions is that major depression is an episodic disorder that expresses itself in spells.20-22 The questions in the DPAX-2 do not assume that depression is episodic. They do, however, elicit more information about functional impairment, and, in this way, implement the criteria for disability introduced in DSM-IV.

The DIS questions are longer and involve more component parts than those in the DPAX-2. Fairly high levels of abstract generalization and acuity of recall are needed. Interviewer comments indicated that some subjects did not understand some DIS questions. Because a higher level of education might make it easier to grasp the meaning of the DIS, we included education in the analysis.

STATISTICAL METHODS

The $\kappa$ statistic was employed for assessing the level of agreement between the results of the different schedules.23 Where the subsample interviewed by the SCID was concerned, estimates of $\kappa$, sensitivity, and specificity were computed after weighting for correspondence to the full sample. We follow the interpretation of Fleiss,24 who suggests that values above 0.40 indicate fair to good agreement and lower values indicate poor agreement. Relationships of sex, age, and education to discordant cases were tested by $\chi^2$ analysis. Responses to individual questions regarding the same demographic factors were tested using logistic regression. Assessment of significance is indicated by 99% confidence intervals.

The odds ratios tended to show that the opposite association with education occurred but was not significant.

QUESTIONS ABOUT ESSENTIAL FEATURES OF DEPRESSION

The 5 questions available for essential features elicited positive responses from different proportions of the sample (Table 4). The DIS question that specifies a 2-year duration of sadness and the DPAX-2 question that uses the idiom of being in poor spirits were answered positively by the fewest number of subjects. On the other hand, close to half of the sample responded positively to the question about feeling low and hopeless. Despite such differences, subjects who answered positively to one question were likely to respond positively to others as illustrated by the fact that 26% ($n = 368$) of the sample responded positively to at least 1 DPAX-2 question and 1 DIS question, and an additional 3% ($n = 42$) responded positively to all 5 questions.

In view of this, the odds ratios tended to show that the individual questions bore similar relationships to sex,
age, and education. For example, positive responses to each question were more characteristic of women than men. There was, however, variability. The DIS question about 2 years of sadness was especially likely to show this sex difference. The DPAX-2 question about wonder worthwhile showed very little difference, suggesting that such a term may be especially meaningful for men.

**SCID STUDY**

The DPAX-2 and DIS information from the subjects selected for the SCID study exhibited approximately the same level of agreement as in the sample as a whole (κ = 0.34 for lifetime depression). Fifty-one subjects were diagnosed by one or the other of the lay-administered methods; of them, 17 were diagnosed by both.

The agreement between the SCID and each of the lay-administered interviews was poor, especially from the perspective of the weighted estimates (Figure 2). Both the DPAX-2 and the DIS had low sensitivity and high specificity. Using the SCID, the clinicians diagnosed about twice as many subjects as having had a depression, mainly MDE, than either of 2 lay-administered methods.

A main purpose in our comparison of the 2 lay-administered interviews was to see if differences would suggest ways in which the content and structure of questions could better achieve the goal of allowing people in the general population to give accurate and reliable information about the manifestations of depression.

In our view, the questions about essential feature work best if they identify all those persons who could conceivably meet other criteria required for a diagnosis. These questions should demonstrate extremely high sensitivity even though they may involve low specificity. The subsequent application of other criteria about duration, associated symptoms, and impairment can remove, step by step, those who do not qualify for a diagnosis, thereby improving specificity with minimal loss of sensitivity.

Survey methods and psychometric studies indicate that reliability is improved when questions are stated simply, when they involve only one idea, and when there are several on the same theme. This suggests that it is valuable to use multiple questions that employ different idioms and phrases for the essential features.

### Table 1. Standardized Current and Lifetime Prevalence of Depression Comparing the DPAX-2 and the Diagnostic Interview Schedule (DIS) in the 1992 Sample (N = 1396)*

<table>
<thead>
<tr>
<th>Method</th>
<th>Current Depression</th>
<th></th>
<th></th>
<th>Lifetime Depression</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age &lt;45 y</td>
<td>Age ≥45 y</td>
<td>Total</td>
<td>Age &lt;45 y</td>
<td>Age ≥45 y</td>
<td>Total</td>
</tr>
<tr>
<td><strong>DPAX-2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>8.2 (1.6)</td>
<td>6.3 (1.2)</td>
<td>7.2 (1.0)</td>
<td>11.0 (1.8)</td>
<td>7.7 (1.3)</td>
<td>9.3 (1.1)</td>
</tr>
<tr>
<td>Male</td>
<td>3.2 (1.1)</td>
<td>5.3 (1.2)</td>
<td>4.2 (0.8)</td>
<td>5.0 (1.4)</td>
<td>6.9 (1.4)</td>
<td>5.9 (1.0)</td>
</tr>
<tr>
<td>Total</td>
<td>5.6 (1.0)</td>
<td>5.8 (0.8)</td>
<td>5.7 (0.6)</td>
<td>8.0 (1.1)</td>
<td>7.3 (1.0)</td>
<td>7.6 (0.7)</td>
</tr>
<tr>
<td><strong>DIS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>9.3 (1.7)</td>
<td>6.4 (1.2)</td>
<td>7.8 (1.0)</td>
<td>17.5 (2.2)</td>
<td>13.9 (1.7)</td>
<td>15.6 (1.4)</td>
</tr>
<tr>
<td>Male</td>
<td>3.7 (1.2)</td>
<td>2.7 (0.9)</td>
<td>3.2 (0.7)</td>
<td>7.0 (1.7)</td>
<td>5.3 (1.2)</td>
<td>6.2 (1.0)</td>
</tr>
<tr>
<td>Total</td>
<td>6.4 (1.0)</td>
<td>4.6 (0.8)</td>
<td>5.5 (0.6)</td>
<td>12.2 (1.4)</td>
<td>9.7 (1.1)</td>
<td>10.9 (0.9)</td>
</tr>
</tbody>
</table>

* The rates were standardized by the direct method to the pooled adult population of Stirling County for 1952, 1970, and 1992 using sex and age as well as 3 districts in which different sampling ratios were used. DPAX indicates version 2 of the Stirling County Study method (DP for depression and AX for anxiety).
The DSM-III and DSM-IV indicate that there are numerous ways of describing depressed mood and anhedonia. Each version indicates that clinicians need to be alert to colloquial variants. In reference to anhedonia, for example, DSM-IV indicates that some individuals complain of feeling “blah.” Through multiple questions using different idioms, it is possible to create a set, so that if one question is unfamiliar to one segment of the population, it will be compensated for by another question that is familiar to that segment.

Our emphasis on multiple questions about essential features stems partly from experience in the Stirling County Study. Our original interview schedule used only the question about poor spirits. At that time and continuing into 1970, this question was answered positively by nearly equal proportions of men and women, thus contributing to our evidence that the prevalence of depression did not seem to differ much by sex. Also, we have shown that if we had continued to use only this one question, it would have been necessary to conclude that the prevalence of depression had decreased. This would have been a mistake, however, because the other indicators of dysphoria increased over time, suggesting that it was response to the idiom of poor spirits that decreased, not the prevalence of depression. Because one of the tasks of psychiatric epidemiology is to document time trends, multiple questions provide a safeguard against the effect of changes in the meaning of specific idioms.

Two different ways of applying criteria for duration were illustrated. The terms 2 weeks and 2 years were embedded in the DIS questions about essential features, while separate and open-ended questions were used by the DPAX-2. There is growing evidence that depression is often a chronic disorder with a fluctuating course rather than a disorder that expresses itself mainly in discrete episodes, as assumed in DSM-III and DSM-IV. If future research comes to reflect this view, interviewing methods that assess duration separately would be better able to adjust than those that build duration into the key questions.

Interpretation about the methodologic differences in assessing associated symptoms points to the fact that the DPAX-2 uses a narrower and more homogeneous definition of depression than the DIS. This difference contributes substantially to disagreement between the methods but does not lead to suggestions about interview improvement, since the DIS approach is more in line with the guidelines in current manuals.

While the DIS deals more adequately with associated symptoms, the DPAX-2 provides more information about impairment. Subjects diagnosed as dysthymic were particularly likely to have not met the DPAX-2 criteria for impairment, suggesting that the present definition of dysthymia may not include sufficient evidence of disturbed functioning to warrant being counted as a disorder.

Distinctive modes of inquiry were employed by the 2 methods. We believe that many of the DIS questions ask subjects to recall the past and to think in complex ways that might be favored by those who have more education. This view is supported by evidence that people of less education, while meeting the requirements of the simpler DPAX-2 procedures, tended not to fulfill the criteria of the DIS. This suggests the need for finding ways to achieve precision by simpler mechanisms.

While several suggestions about ways in which interview schedules might be improved emerged from this
analysis, experience should not be abandoned in favor of entirely new directions. One reason for satisfaction with progress stems from evidence about the reliability of responses. Even across the differences in words and sentence structure illustrated by the DPAX-2 and DIS, many people indicated through the consistency of their responses that they were listening carefully and understanding the questions. Another reason for methodologic optimism is that DSM-III gave an unprecedented platform for schedule construction by describing the structure of a diagnosis as composed of essential features, duration, and associated symptoms, to which DSM-IV has added the components of distress and disability. The DIS implemented this conception, and its heir, the Composite International Diagnostic Interview, maintains it while implemented this conception, and its heir, the Composite International Diagnostic Interview, maintains it while added the components of distress and disability. The DIS, as measured by the statistic) and by sensitivity and specificity showing unweighted results as well as estimates weighted so that the subsample interviewed by SCID reflects the whole sample. CI indicates confidence interval; plus sign, positive for depression; and minus sign, negative for depression.

However, whether the SCID assessments are more valid than the others remains a question, since more evidence of the SCID’s reliability than validity has thus far been accumulated. It will be useful if the results of SCID can be evaluated using the “LEAD standard” (L indicates longitudinal; E, experts; and AD, all data), which draws on longitudinal information evaluated by clinical experts who base their judgments on all data accumulated from various sources. The LEAD standard emphasizes that longitudinal information offers an important means of assessing the degree to which disorders in communities are clinically significant. In this regard, a 16-year follow-up study of 36 women and 24 men diagnosed as depressed by the DPAX-1 in our first study indicated that course and outcome were marked by continuing morbidity and premature mortality. Depressed men, none of whom suffered from alcohol abuse, exhibited twice the number of expected deaths. While deaths among depressed women were not appreciably above expectation, one depressed woman did commit suicide. Of the 24 women and 8 men who survived

### Table 4. Demographic Factors in Relation to Positive Responses to Questions About Essential Features of Depression Employed in the DPAX-2 and the Diagnostic Interview Schedule (DIS)*

<table>
<thead>
<tr>
<th>Questions</th>
<th>Positive Responses, %</th>
<th>Sex (Men as Reference)</th>
<th>Age (≥45 y as Reference)</th>
<th>Education (Grade ≥11 as Reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>From DPAX-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor spirits</td>
<td>13.5</td>
<td>1.5 (1.1-2.1)</td>
<td>1.9 (1.4-2.6)</td>
<td>1.9 (1.3-2.6)</td>
</tr>
<tr>
<td>Wonder worthwhile</td>
<td>44.0</td>
<td>1.2 (1.0-1.5)</td>
<td>1.5 (1.2-1.9)</td>
<td>1.4 (1.1-1.8)</td>
</tr>
<tr>
<td>Low and hopeless</td>
<td>46.7</td>
<td>1.8 (1.5-2.3)</td>
<td>1.3 (1.1-1.7)</td>
<td>1.3 (1.0-1.6)</td>
</tr>
<tr>
<td>From DIS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Weeks sad</td>
<td>32.2</td>
<td>1.7 (1.3-2.1)</td>
<td>1.0 (0.8-1.3)</td>
<td>1.0 (0.8-1.3)</td>
</tr>
<tr>
<td>2 Years sad</td>
<td>11.5</td>
<td>3.0 (2.1-4.4)</td>
<td>0.9 (0.6-1.2)</td>
<td>1.8 (1.3-2.6)</td>
</tr>
</tbody>
</table>

* Percentages and odds ratios are based on the responses given by the 1396 subjects of the 1992 sample. DPAX-2 indicates version 2 of the Stirling County Study method (DP for depression and AX for anxiety); CIs, confidence intervals.

Figure 2. Agreement and disagreement about lifetime depression between the Structured Clinical Interview for DSM-III-R (SCID) and 2 methods of lay-administered interviews (version 2 of the Stirling County Study’s customary method, the DPAX [DP for depression and AX for anxiety]); and the Diagnostic Interview Schedule [DIS] as measured by the statistic and by sensitivity and specificity showing unweighted results as well as estimates weighted so that the subsample interviewed by SCID reflects the whole sample. CI indicates confidence interval; plus sign, positive for depression; and minus sign, negative for depression.
and were reinterviewed, 78% (n = 25) remained chronically or recurrently ill with either a depression or anxiety disorder, with more women than men crossing over to anxiety.14 The severity of depression among men was emphasized by the fact that, at the end of 16 years, we found that 83% (n = 20) of them either were dead or had been chronically or recurrently depressed with persistent impairment throughout the period, while the same outcome applied to only 42% (n = 15) of depressed women.16

Our methods of diagnosis were intended to identify people who had a need for psychiatric attention. Information about the degree to which these depressed people did in fact seek attention was limited to those who were alive at the end of the follow-up interval. This perspective indicated that 83% (n = 20) of the depressed women had sought attention for their emotional problems from a physician, usually a general physician.15 Although the outcome was poorer for depressed men, only 37% (n = 3) of them had sought help through such sources.

Such long-term information indicates that the depressions identified by the lay-administered DPAX-1 method were not “normal and transient responses.” Rather, the evidence betokens a pathologic process. Spitzer reported the pathologic process as bearing a “character of danger” for survival and well-being, a character that has been borne out in the evidence about our “character of danger” for survival and well-being, a character that has been borne out in the evidence about our

Accepted for publication August 19, 1999.

This work was supported by grant MH39576 from the National Institute of Mental Health, Rockville, Md (Dr Murphy).

We thank James Leavitt, MD, Martha Shenton, PhD, Carolyn Turvey, PhD, David Vogel, PhD, and John Worthington, MD, for administering the SCID interviews; Constantine Daskalakis, DSc, for statistical help; Patricia Merritt, Barbara Burns, and Ellen Krystofik for field management and data processing; and Barbara Burns and Ellen Krystofik for help with manuscript preparation.

Reprints: Jane M. Murphy, PhD, Department of Psychiatry, Massachusetts General Hospital, Room 9155, 149 13th St, Charlestown, MA 02129-2000 (e-mail: murphy.jane@mgh.harvard.edu).

REFERENCES


