ORIGINAL ARTICLE

A Comparison of Self-report and Clinical Diagnostic Interviews for Depression

Diagnostic Interview Schedule and Schedules for Clinical Assessment in Neuropsychiatry in the Baltimore Epidemiologic Catchment Area Follow-up

William W. Eaton, PhD; Karen Neufeld, MD; Li-Shiun Chen, MD; Guojun Cai, MD

Background: The field of psychiatric epidemiology continues to employ self-report instruments, but the low degree of agreement between diagnoses achieved using these instruments vs that achieved by psychiatrists in the clinical modality threatens the credibility of the results.

Methods: In the Baltimore Epidemiologic Catchment Area follow-up, 349 individuals who had a Diagnostic Interview Schedule (DIS) interview were blindly examined by psychiatrists using the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). Comparisons were made at the level of diagnosis, syndrome, and *DSM-IV* symptom group. Indexes of agreement were computed and characteristics of discrepant cases were identified.

Results: Agreement on diagnosis of major depressive disorder was only fair ($\kappa = 0.20$), with the DIS missing many cases judged to meet criteria for diagnosis using the SCAN (29% sensitivity). A major source of discrepancy was respondents with false-negative diagnoses who repeatedly failed to report DIS symptoms attributed to life crises or medical conditions. Older age, male sex, and lower impairment were associated with underdetection by the DIS, using logistic regression analysis. In spite of the diagnostic discrepancy, there was substantial correlation in numbers of symptom groups in the 2 modalities ($r = 0.49$). Agreement was highest (about 55% sensitivity and 90% specificity) when both the SCAN and DIS thresholds were set at the level of depression syndrome instead of diagnosis.

Conclusions: Weak agreement at the level of diagnosis continues to threaten the credibility of estimates of prevalence of specific disorders. A bias toward underreporting, as well as stronger agreement at the level of the depression syndrome and on ordinal measures of depressive symptoms, suggests that associations with risk factors are conservative.

Arch Gen Psychiatry. 2000;57:217-222

From the Department of Mental Hygiene, School of Hygiene and Public Health (Drs Eaton and Chen), and the Department of Psychiatry, School of Medicine (Dr Neufeld), Johns Hopkins University, Baltimore, Md; and the Center for Health Development, World Health Organization, Kobe, Japan (Dr Cai).

AN IMPORTANT justification for the Epidemiologic Catchment Area (ECA) program, and the "third generation" of psychiatric epidemiology ushered in by the program, was the purported ability to link results from epidemiologic research to results from research conducted in laboratories and clinics.[1-3] The link was made possible by the structure of the Diagnostic Interview Schedule (DIS),[4] which was designed to generate information of the type obtained in a psychiatric interview. A potential weakness of the DIS was its reliance on the judgments and insights of the respondent instead of a psychiatrist. Validation studies in medical settings have shown good agreement between the DIS and a clinical interview[4]; however, validation studies conducted in the context of household surveys show weaker agreement.[5,6]

Much of the database in psychiatric epidemiology originates from household interviews conducted with self-report interviews, such as the DIS or its descendant, the Composite International Diagnostic Interview.[7] The credibility of these surveys has been challenged because of discrepancies with psychiatric diagnoses and because of anomalies in the data, such as decline in lifetime prevalence of disorders with age.[8,9] Some feel that the DIS-style interviews capture a wider range of trivial or less severe symptoms and disorders than a psychiatrist would, and that the high estimates of prevalence and incidence are biased.[10]

See also pages 209, 223, 227, and 230

In 1993 through 1996, follow-up evaluations were conducted on the cohort of 3481 individuals at the Baltimore ECA site.[11] The follow-up evaluations included a clinical examination of a

SUBJECTS AND METHODS

SAMPLING AND RESPONSE

The target population in 1981 was the 175 211 household residents of east Baltimore. In 1981, 4238 individuals were probabilistically designated for interview in the first wave; 3481 (82%) completed the interviews. In 1993, these 3481 were the target for interviewing, carried out from 1993 through 1996 (median follow-up period, 12.6 years). The original sample overrepresented the elderly somewhat, and there were 848 respondents (24%) who died before the follow-up interview. More than 73% of those alive were interviewed.[14] Depressive disorder was not significantly related to mortality, loss of contact, or refusal to be interviewed.[14] Demographically, the 3481 respondents in 1981 were 62% female, 26% older than 64 years, and 34% African American; the respondents participating in the follow-up interviews were 63% female, 14% older than 64 years, and 33% African American.

DIS INTERVIEW

The interview was designed to be as close as possible to that conducted in 1981. Interviews took place in the respondents' household, using the DIS, version III, revised.[15] Interviewers were typical survey interviewers (ie, mostly middle-aged women with good verbal skills, but there were no requirements for educational background or clinical experience). They received 2 weeks of training. Diagnostic algorithms for *DSM-III-R* (created by the originators of the DIS) were used. Subjects discussed and signed informed-consent documents both for the household phase of the research and for the office-visit phase discussed below, in conformity with requirements of the Johns Hopkins University School of Public Health Institutional Review Board, Baltimore, Md.

SCAN INTERVIEW

Respondents who, on the basis of their baseline and follow-up responses, seemed to be potential new cases with any of 7 disorders (major depressive disorder, panic disorder, agoraphobia, social phobia, obsessive-compulsive disorder, alcohol use disorder, and cognitive impairment) were invited to a local office for an examination by a psychiatrist and a series of personality, risk factor, and neuropsychological tests.[13] A small random sample of all respondents was also invited to the office (41 subjects who completed office visits and who did not meet other selection criteria). The total number invited was 585, of whom 349 (60%) completed the SCAN. Demographically, the office visit sample was 64% female, 7% older than 64 years, and 36% African American. Presumably, the fact that selection into the office visit sample weighted heavily those with new-onset depression led to the underrepresentation of the elderly. The median time between DIS and SCAN interviews was 113 days, with 57 and 219 days as the 25th and 75th percentiles, respectively.

The SCAN incorporates the 10th edition of the Present State Examination[16] and retains many of its features, such as rating scales with defined thresholds, a glossary of definitions, and a semistructured approach. The reliability of the SCAN is good.[17] For the Baltimore ECA follow-up, several items were added to the SCAN, version 1.0, consistent with our interest in natural history, and to ensure that *DSM-III-R* diagnoses could be made. Some of these additions were incorporated into the second edition of the SCAN.[18] In the area of depression, the following items were added: social withdrawal, increase in appetite, age at first onset of dysphoria or anhedonia, age at first onset of depressive delusions, organic cause of symptoms about thinking and concentration, and age at first onset of sleep problems. A computer algorithm was written for the *DSM-III-R* diagnosis, adapting as much as possible the World Health Organization SCAN algorithm previously written for the *International Classification of Diseases, 10th Revision*.

The SCAN was administered by board-certified psychiatrists who had completed a 1-week SCAN training course at the Johns Hopkins World Health Organization SCAN training center: one psychiatrist was trained at the London World Health Organization SCAN training center. The psychiatrists were kept blind to the results of the DIS until the entire SCAN had been administered. After the examination was concluded, a 2- or 3-page summary was dictated for each subject in a standard format to fill in the clinical picture, ensure inclusion of important idiosyncratic details, and explain possible disagreements with the DIS as well as diagnostic judgments that might not seem to be justified on the evidence when taken on an item-by-item basis. For the first 20 interviews of each psychiatrist, cases were discussed in detail at a group conference with 2 or more other interviewing psychiatrists. After the first 20, full-scale case conferences were scheduled only for difficult cases when requested by the interviewing psychiatrist; all other cases were discussed in a conference with a minimum of 1 other interviewing psychiatrist.

STATISTICAL ANALYSES

Agreement between DIS and SCAN was measured using the κ statistic for diagnostic assessments and the correlation coefficient for interval assessments of number of symptoms. The κ value weights disagreement between adjacent categories as equivalent to disagreement between very distant categories; in effect, this ignores some of the association revealed by the correlation. A κ statistic using linear weights $(1-[i-j]/[k-1]$, where i and j index the rows and columns of the table and k is the maximum number of categories) is computed, as well as a weighted κ value using squared linear weights; in effect, this weights disagreement in adjacent categories minimally. Sensitivity and specificity are computed, temporarily assuming that the SCAN is the criterion standard. A logistic regression focused on the distinction between false-negative diagnoses (coded as 1) and true-positive diagnoses (coded as 0). Probabilities of selection, available at each stage of sampling from the 1981 designation to the invitation and completion of the office visit, permit estimation of weights for inference to the survivors of the 1981 cohort population. These weights are not used, however, since the interest is analytic, not descriptive, and since the target population of survivors is of less intrinsic interest than the household-residing population in 1981. Weighted analyses (data not shown) did not differ in strength or significance from those shown herein.

subsample of respondents by a psychiatrist using the Schedules for Clinical Assessment in Neuropsychiatry (SCAN),[12] conducted after the DIS household interview. A previous article on the incidence of panic disorder included some details of the SCAN follow-up method.[13] The present article compares data on depression obtained by interviewers using the DIS and by psychiatrists using the SCAN, including agreement at the level of symptom and syndrome as well as diagnosis.

RESULTS

MEASURES OF AGREEMENT

About one fifth of the respondents in the office-visit phase of the study received a lifetime diagnosis of depressive disorder on either the DIS or the SCAN (78/349) (**Table 1**). Using the diagnostic algorithms and lifetime diagnosis, the level of agreement between the 2 methods is not strong (Table 1). The DIS recorded a positive diagnosis in 23 of the 78 cases diagnosed by the psychiatrist (29% sensitivity). Two hundred sixty of the 271 SCAN-negative cases were also found to be negative using the DIS (96% specificity).

One might expect that agreement would be greater for cases with more symptoms. **Table 2** displays agreement as to whether the SCAN or the DIS obtained a positive rating for the presence in the lifetime of a symptom in the 9 *DSM-III-R* symptom groups (dysphoria, anhedonia, appetite/weight, sleep, psychomotor, fatigue, guilt, concentration, and suicidal ideation). For both the DIS and the SCAN, the symptoms occurred during a presumed "worst" episode, with dysphoria or anhedonia present. The boxes along the diagonal represent exact agreement on the number of symptom groups (not identical to agreement on every single symptom). Outside the agreement on symptoms in no groups at all (n = 119), there are only 13 subjects with this level of agreement. The gray areas represent subjects for whom there were discrepancies greater than 5 symptom groups. Disagreement is greater at the lower thresholds, primarily because the SCAN obtains symptoms in 5 or more groups, whereas the DIS records none (45 subjects in the gray area in the top row of Table 2), including 8 cases with symptoms in 7 groups and 3 with symptoms in 8 groups. These strong false negatives dominate the table, contrasting with only 6 strong false positives at the bottom left of the table.

There is substantial association between the DIS and the SCAN. For example, the Pearson correlation for number of symptom groups is 0.49, but the chance-corrected measure of agreement (κ statistic) for Table 2 is weak, with a value of 0.20. The linear-weighted κ value is 0.31, and the κ value with squared weights approaches the correlation, with a value of 0.43.

The measurement qualities of the DIS and SCAN are related to the threshold for judging a case as positive. **Figure 1** shows the sensitivity and specificity for each of the 9 thresholds that could be applied to DIS data. The criterion closest to diagnosis is the distinction between 0 to 4 vs 5 to 9 symptom groups, which shows a sensitivity of about 30% and specificity of about 95%, as in Table 1. As expected, the sensitivity declines and the specificity rises as the DIS threshold is raised. However, the specificity is hardly affected at all by this threshold change and, consistent with the high proportion of strong false-negative diagnoses, the sensitivity is not above 50% even when the threshold is at only 1 symptom group. Changes in the SCAN threshold hardly affect these relationships. If the SCAN diagnosis is broadened to a threshold of 0 to 3 vs 4 to 9 symptom groups, the sensitivity rises by about 10%; if the SCAN diagnosis is narrowed to a threshold of 0 to 6 vs 7 to 9 symptom groups, the sensitivity drops by about 5%; the specificity is almost unchanged (data not shown). The concept of depression syndrome—an episode of dysphoria or anhedonia accompanied by symptoms in 2 or more *DSM-III-R* symptom groups (ie, symptoms in 3 groups)—generates the highest agreement. This concept does not use diagnostic exclusions and does not require that the full criteria for diagnosis ever have been met in the lifetime. The threshold of 0 to 3 vs 4 or more symptom groups, applied to both the SCAN and the DIS, yields a sensitivity of about 55% and specificity of about 90% (data not shown).

Table 1. Discrepancy Between the DIS and SCAN for the Lifetime Occurrence of Depressive Disorder in the Baltimore ECA Follow-up*

	Psychiatrist Using SCAN		
Interviewer Using DIS	**Never a Case**	**Positive Diagnosis**	**Total**
Never a case	260	55	**315**
Positive diagnosis	11	23	**34**
Total	**271**	**78**	**349**

**DIS indicates Diagnostic Interview Schedule; SCAN, Schedules for Clinical Assessment in Neuropsychiatry; and ECA, Epidemiologic Catchment Area.*

ANALYSIS OF DISCREPANCIES

The DIS responses were examined for 11 individuals with 7 or 8 symptom groups by the SCAN but 0 symptom groups by the DIS (top right of Table 2). All but 2 of these individuals reported some depressive symptoms over their lifetimes. Five of the 11 subjects attributed all or part of their depressive symptoms to medication, drugs or alcohol, or physical illness or injury. Five other subjects had lifetime symptoms in fewer than 3 symptom groups, 1 subject had neither dysphoria nor anhedonia, and 2 subjects reported that the symptoms did not cluster together in 1 period. As a result, they were all excluded from questions regarding symptoms during the worst episode, which was required for inclusion in the symptom group algorithm used in Table 2. This reveals an important difference between the DIS and the SCAN, in that the DIS assessment of episode is highly structured and does not facilitate revision during the interview, whereas the SCAN interviewer can record symptoms present without a rigid definition of *episode*, deciding during the interview process when the worst episode might have occurred. This difference has little effect on diagnostic discrepancy, but leads to an underestimate of the correlation of 0.49 for number of symptom groups.

Table 2. Discrepancy Between the DIS and SCAN in the Number of *DSM-IV* Symptom Groups Reported for the Worst Episode of Depression in the Baltimore ECA Follow-up*

No. of *DSM-IV* Symptom Groups by Interviewer Using DIS	No. of *DSM-IV* Symptom Groups by Psychiatrist Using SCAN, No. of Subjects										
	0	1	2	3	4	5	6	7	8	9	Total
0	119	51	28	12	11	22	12	8	3	0	**266**
1	0	1	2	3	0	0	0	0	0	0	**6**
2	2	3	0	0	0	1	0	0	0	0	**6**
3	0	2	2	2	1	0	1	0	1	0	**9**
4	1	4	4	2	0	4	3	2	2	1	**23**
5	1	1	0	1	1	1	2	3	4	0	**14**
6	0	1	1	1	0	1	3	2	0	0	**9**
7	0	1	2	0	1	0	0	6	0	1	**11**
8	0	0	0	0	2	0	0	0	0	0	**2**
9	0	0	0	0	1	0	0	0	2	0	**3**
Total	**123**	**64**	**39**	**21**	**17**	**29**	**21**	**21**	**12**	**2**	**349**

**Gray areas indicate subjects with discrepancies greater than 5 symptom groups. The boxes along the diagonal represent exact agreement on the number of symptom groups. DIS indicates Diagnostic Interview Schedule; SCAN, Schedules for Clinical Assessment in Neuropsychiatry; and ECA, Epidemiologic Catchment Area.*

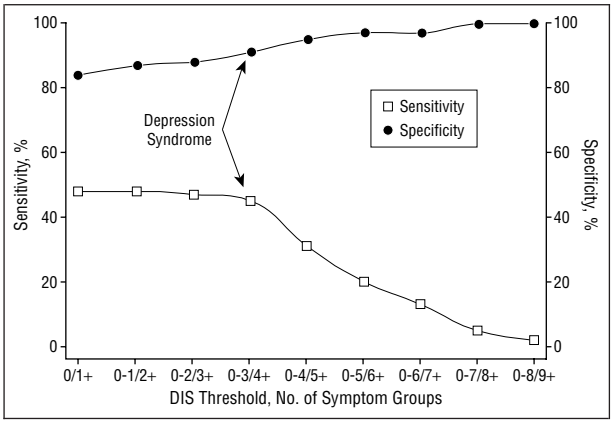

Figure 1. *The effect of thresholds on diagnostic agreement for the Baltimore Epidemiologic Catchment Area follow-up evaluations.*

Table 3. Predictors of False-Negative Depressive Disorder Using the Diagnostic Interview Schedule*

Variable	Odds Ratio (95% CI) by Logistic Regression	
	Unadjusted	Adjusted†
Age, y		
18-29	1.00	1.00
30-44	1.58 (0.56-4.46)	1.06 (0.26-4.21)
≥45	6.81 (0.78-59.1)	6.12 (0.55-67.8)
Sex		
Female	1.00	1.00
Male	2.73 (0.71-10.5)	3.41 (0.61-19.1)
Race		
Nonblack	1.00	1.00
Black	1.64 (0.58-4.64)	2.80 (0.62-12.7)
Recency of episode		
≤1 y	1.00	1.00
>1 y	2.12 (0.72-6.20)	3.69 (0.93-14.6)
Impairment		
Little	5.32 (0.57-49.2)	1.48 (0.11-20.5)
Moderate	1.72 (0.61-4.88)	0.78 (0.17-3.51)
Severe	1.00	1.00
Symptoms, No.		
7-9	1.00	1.00
5-6	5.57 (1.93-16.1)	4.42 (1.17-16.7)

**For 23 true positives compared with 55 false negatives. CI indicates confidence interval.*
†*Adjusted for all other variables in the model.*

The psychiatrist code books and dictations of SCAN interviews were reviewed to learn more about the characteristics of these same 11 DIS-negative and SCAN-positive respondents. Four of these SCAN-positive and DIS-negative records met the criteria for major depressive episodes in the midst of stressful life events. For example, a 44-year-old woman who reported no symptoms on the DIS gave a history of episodes of depression after the deaths of 2 family members to the psychiatrist. The first episode lasted 7 months, and was characterized by vegetative symptoms and suicidal preoccupation, and resulted in job loss. Despite the respondent's attribution of her symptoms to grief, they were formulated by the psychiatrist as major depressive episodes on the basis of their severity, character, and duration. Another 3 of the SCAN-positive and DIS-negative cases met the criteria for major depressive episodes during heavy drug or alcohol use, but the psychiatrist chose not to exclude the diagnosis of major depression. Two more cases revealed episodes in the midst of other physical disorders. For example, a 51-year-old man attributed depressive symptoms to his chronic back pain, which he stated "is more disabling than my doctors believe it should be." Two of 11 SCAN cases with no DIS symptoms did not have any concurrent events or illnesses but were felt to have met the criteria for a major depressive episode in the midst of a bipolar type I or type II disorder.

A second group of 6 records with SCAN interviews reporting few if any depressive symptoms and 5, 6, 7, or 9 symptoms reported using the DIS were reviewed (bottom left of Table 2). These SCAN-negative and DIS-positive cases revealed 4 interviews in which the psychiatrist found depressive symptoms, but not at a severity consistent with SCAN criteria for major depressive episodes.

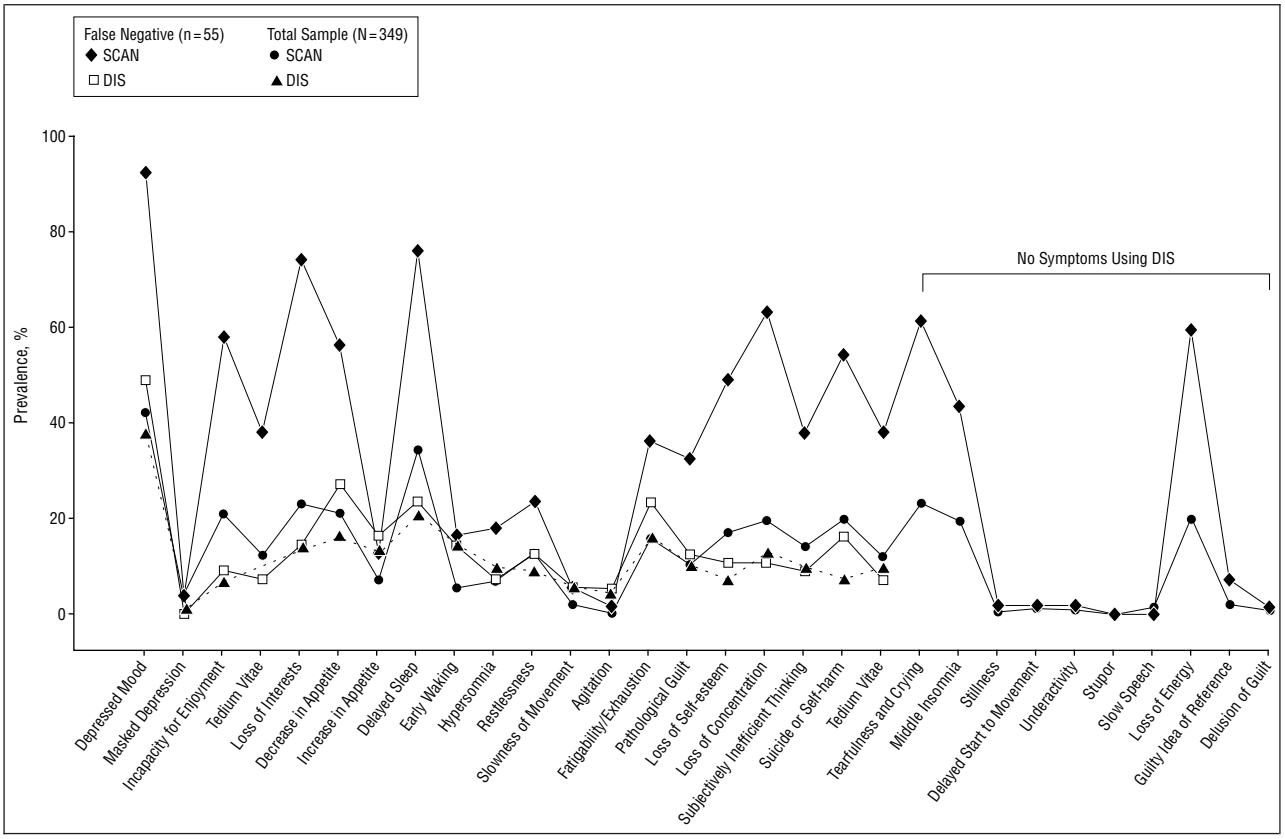

Figure 2. *Prevalence of symptoms using the Schedule for Clinical Assessment in Neuropsychiatry (SCAN) and Diagnostic Interview Schedule (DIS) for the Baltimore Epidemiologic Catchment Area follow-up evaluations.*

The SCAN-positive and DIS-negative discrepant cases were individuals who had depressive symptoms but did not report them at all in the DIS because they were attributed to some nonpsychiatric occurrence or medical condition, or whose reports of symptoms were discounted because of the DIS probe structure. The SCAN-negative and DIS-positive cases were far fewer and revealed individuals who actually reported symptoms that were rated as less severe in the psychiatrist interview. The use of a single threshold exaggerates discrepancy by placing individuals with similar but not identical symptom profiles into an all-or-nothing discrepant category.

What type of respondent is associated with diagnostic discrepancy between DIS and SCAN? We examined predictors for DIS underdetection, including age, sex, race, education, recency of depressive episode, depression-related impairment, number of comorbid DIS disorders, and number of depressive symptoms (**Table 3**). Effects of education and comorbidity were small and not significant. Individuals who were depressed with 5 or 6 depressive symptom groups, compared with those with 7 to 9 symptom groups, were more likely not to be detected when using the DIS assessment, with an odds ratio of 4.42 (95% confidence interval, 1.17-16.7; $P = .03$). Older age, being male, having a less recent episode, and having less depression-related impairment were associated with DIS underdetection, although these odds ratios were not significant because of the small sample size.

What type of symptom is likely be associated with diagnostic discrepancy between the DIS and the SCAN? **Figure 2** shows the symptom profiles for SCAN and DIS for the entire sample of 349 subjects and for the 55 false negatives. The SCAN includes 10 symptoms relevant to depression that are absent from the DIS, shown in Figure 2. Except for loss of energy, these latter symptoms have a prevalence below 10%. In the sample as a whole, certain symptoms are recorded much more frequently in the SCAN than in the DIS, including incapacity for enjoyment, loss of interest, delayed sleep, loss of self-esteem, loss of concentration, and suicidal ideation. The same symptoms are closely linked to those with false-negative diagnoses, being much more prevalent in the SCAN record than in the DIS record for those 55 subjects. Other important differences in symptom prevalence in those with false-negative diagnoses occur for tedium vitae and subjectively inefficient thinking. In contrast, there is good agreement for increased appetite, early waking, hypersomnia, restlessness, slowness of movement, and agitation and moderate agreement for fatigue.

COMMENT

These analyses have important limitations. The household sample suffered considerable attrition, and it represents only one area of a single city. The design for sampling respondents for the SCAN was toward new cases, not prevalent cases. For these reasons it is difficult to generalize these results to other ECA sites or to other epidemiologic studies. Although agreement was assessed for lifetime occurrence, there was a substantial

interval between the DIS and the SCAN interview. The sample size limited our ability to study current symptoms and diagnoses, since these were much rarer than lifetime symptoms and diagnoses; however, the analysis of recency suggests that agreement would have been higher for current than for lifetime diagnosis. Use of different assessment procedures necessarily entails different sets of symptoms, as well as different diagnostic algorithms; this may contribute information and criterion variation to the analysis and may raise the amount of discrepancy in diagnosis. Most of these problems lead to an underestimate of agreement between the DIS and SCAN.

It is difficult to imagine progress in the field of psychiatric epidemiology without instruments such as the DIS and Composite International Diagnostic Interview as part of the measurement armamentarium. It is unlikely that the highly structured self-report modality will ever be satisfactory for disorders such as schizophrenia or bipolar disorder, in which lack of insight precludes relying heavily on the subject's judgment as to the presence or absence of a symptom, or the impairment it may generate.[19] For other disorders, however, in which subjects have insight into their own mental life, the survey modality has strong appeal because of its cost-effectiveness and the large biases in studying cases from clinical populations. These findings suggest that the potential for self-report instruments is greatest when the results are not strictly dependent on the threshold for the presence or absence of a specific diagnosis. Determinations of prevalence of disorders (a focus of descriptive epidemiology) are dependent on the threshold. However, in analytic studies, valid associations will be detected with risk factors for disorders and syndromes and their consequences, even in the presence of diagnostic discrepancies with other modalities of measurement. This is especially true when there is a high threshold for the diagnosis (as apparently is the situation with the DIS), or when ordinal or interval measures of intensity or the number of symptoms are used. For example, because of the high specificity of the DIS for depression and panic,[13] conclusions about risk-factor associations are likely to be conservative (ie, the false-negative cases will dilute the degree of association by mixing with the noncases). Since the noncases are the majority in population-based surveys, the effects of dilution will be moderate or small.[20] The value of prevalence studies may be overrated in psychiatric epidemiology, as compared with other areas of epidemiology.[21] In our opinion, the field will benefit by decreasing the focus on descriptive epidemiology of specific diagnoses and increasing the focus on analytic epidemiology of syndromes. Failure to find high diagnostic agreement between the SCAN, developed for clinical use, and the DIS, developed for epidemiologic purposes, is not dissimilar from failures in other areas of medicine. However, diagnostic discrepancy does not inevitably vitiate the ability to link etiologic clues from epidemiologic research with etiologic theories arising from laboratory and clinical research.

Accepted for publication June 16, 1999.

This research was supported by grant MH47447 from the National Institute of Mental Health, Rockville, Md (Dr Eaton).

Corresponding author: William W. Eaton, PhD, Department of Mental Hygiene, Johns Hopkins University, 624 N Broadway, Room 880, Baltimore, MD 21205-1999.

REFERENCES

1. Eaton WW, Regier DA, Locke BZ, Taube CA. The NIMH Epidemiologic Catchment Area program. In: Wing JK, Bebbington P, Robins LN, ed. *What Is a Case? The Problem of Definition in Psychiatric Community Surveys*. London, England: Grant McIntyre Ltd; 1981:99-106.
2. Regier DA, Myers JK, Kramer M, Robins LN, Blazer DG, Hough RL, Eaton WW, Locke BZ. The NIMH Epidemiologic Catchment Area (ECA) program: historical context, major objectives, and study population characteristics. *Arch Gen Psychiatry*. 1984;41:934-941.
3. Eaton WW, Kessler LG, eds. *Epidemiologic Field Methods in Psychiatry: The NIMH Epidemiologic Catchment Area Program*. Orlando, Fla: Academic Press Inc; 1985.
4. Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics, and validity. *Arch Gen Psychiatry*. 1981;38:381-389.
5. Helzer JE, Robins LN, McEvoy LT, Spitznagel EL, Stoltzman RK, Farmer A, Brockington IF. A comparison of clinical and Diagnostic Interview Schedule diagnoses: reexamination of lay-interviewed cases in the general population. *Arch Gen Psychiatry*. 1985;42:657-666.
6. Anthony JC, Folstein MF, Romanoski AJ, Von Korff MR, Nestadt GR, Chahal R, Merchant A, Brown CH, Shapiro S, Kramer M, Gruenberg EM. Comparison of the lay Diagnostic Interview Schedule and a standardized psychiatric diagnosis: experience in eastern Baltimore. *Arch Gen Psychiatry*. 1985;42:667-675.
7. Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, Farmer A, Jablenski A, Pickens R, Regier DA, Sartorius N, Towle LH. The Composite International Diagnostic Interview: an epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Arch Gen Psychiatry*. 1988;45:1069-1077.
8. Parker G. Are the lifetime prevalence estimates in the ECA study accurate? *Psychol Med*. 1987;17:275-282.
9. Regier DA, Kaelber CT, Rae DS, Farmer ME, Knauper B, Kessler RC, Norquist GS. Limitations of diagnostic criteria and assessment instruments for mental disorders: implications for research and policy. *Arch Gen Psychiatry*. 1998;55:109-115.
10. Frances A. Problems in defining clinical significance in epidemiological studies. *Arch Gen Psychiatry*. 1998;55:119.
11. Eaton WW, Anthony JC, Gallo J, Cai G, Tien A, Romanoski A, Lyketsos C, Chen LS. Natural history of Diagnostic Interview Schedule/*DSM-IV* major depression: the Baltimore Epidemiologic Catchment Area follow-up. *Arch Gen Psychiatry*. 1997; 54:993-999.
12. Wing JK, Babor T, Brugha T, Burke J, Cooper JE, Giel R, Jablenski A, Regier D, Sartorius N. SCAN: Schedules for Clinical Assessment in Neuropsychiatry. *Arch Gen Psychiatry*. 1990;47:589-593.
13. Eaton WW, Anthony JC, Romanoski A, Tien A, Gallo J, Cai G, Neufeld K, Schlaepfer T, Laugharne J, Chen LS. Onset and recovery from panic disorder in the Baltimore Epidemiologic Catchment Area follow-up. *Br J Psychiatry*. 1998;173: 501-507.
14. Badawi MA, Eaton WW, Myllyluoma J, Weimer L, Gallo JJ. Psychopathology and attrition in the Baltimore ECA 15-year follow-up 1981-1996. *Soc Psychiatry Psychiatr Epidemiol*. 1999;34:91-98.
15. Robins L, Helzer J, Cottler L, Goldring E. *NIMH Diagnostic Interview Schedule, Version III Revised (DIS-III-R)*. St Louis, Mo: Washington University; 1989.
16. Wing J, Cooper JE, Sartorius N. *The Description and Classification of Psychiatric Symptoms: An Instruction Manual for the PSE and CATEGO System*. London, England: Cambridge University Press; 1974.
17. Tomov T, Nikolov V. Reliability of SCAN categories and scores: results of the field trials. In: Stefanis CN, Rabavilas AD, Soldatos CR, eds. *Vol 1: Classification and Psychopathology, Child Psychiatry, Substance Use: Proceedings of the VIII World Congress of Psychiatry, Athens, 12-19 October 1989*. Amsterdam, the Netherlands: Excerpta Medica; 1990.
18. World Health Organization. *SCAN: Schedules for Clinical Assessment in Neuropsychiatry, Version 2.0*. Geneva, Switzerland: Psychiatric Publishers International/ American Psychiatric Press Inc; 1993-1994.
19. Eaton WW, Romanoski A, Anthony JC, Nestadt G. Screening for psychosis in the general population with a self-report interview. *J Nerv Ment Dis*. 1991;179: 689-693.
20. Kelsey HL, Thompson WD, Evans AS. *Methods in Observational Epidemiology*. New York, NY: Oxford University Press; 1986.
21. Spitzer RL. Diagnosis and need for treatment are not the same. *Arch Gen Psychiatry*. 1998;55:120.