Using Questionnaires to Screen for Psychiatric Disorders: A Comment on a Study of Screening for Bipolar Disorder in the Community

Mark Zimmerman, M.D.; Michael A. Posternak, M.D.;
Iwona Chelminski, Ph.D.; and David A. Solomon, M.D.

In a recent issue of this journal, Hirschfeld and colleagues[1] reported the results of a large epidemiologic study (more than 85,000 participants) of bipolar disorder based on the Mood Disorder Questionnaire (MDQ), a self-administered screening scale for mania/hypomania. The article followed the publication of 2 other reports describing the diagnostic properties of the MDQ in psychiatric patients and a subset of the subjects in the epidemiologic study.[2,3] In the psychiatric patient sample,[2] the MDQ had a sensitivity of 73% and a specificity of 90%, and the respective diagnostic properties in the community sample were 28% and 97%. On the basis of these data, Hirschfeld et al. concluded that the MDQ had "generally good sensitivity" and therefore could be used as a case-finding instrument in their epidemiologic study.

Two important conclusions were drawn from the epidemiologic study.[1] First, Hirschfeld and colleagues suggested "that the actual prevalence of (bipolar disorder) may be higher than had been previously thought."[1(p57)] Second, they suggested that many patients with bipolar disorder may be misdiagnosed with nonbipolar (unipolar) disorder. Hirschfeld et al. advised clinicians to routinely assess for a history of manic/hypomanic symptoms in depressed patients because their presence has treatment and prognostic implications. These findings, thus, have potentially significant public health implications. We believe, however, that a critical examination of the diagnostic properties of the MDQ would suggest that the conclusions drawn from the epidemiologic study may not be justified.

The purpose of the present commentary is to review principles of diagnostic screening and apply them to the studies of the MDQ. We begin by reviewing the goals and purposes of screening. Then, we review the statistics of diagnostic testing. Finally, we critically examine the performance of the MDQ.

PURPOSE OF DIAGNOSTIC SCREENING

The primary reason for screening is to modify the natural history of a disorder by identifying it as early as possible with tests that are usually minimally invasive and relatively inexpensive.[4] The performances of screening tests are compared against a diagnostic "gold standard." For psychiatric disorders, standardized structured diagnostic interviews have been held up as a diagnostic gold standard, albeit an imperfect one. In most contemporary research studies of psychiatric phenomenology, family history, epidemiology, or biology, diagnoses are based on interview schedules such as the Structured Clinical Interview for DSM-IV (SCID).[5]

In clinical practice, these interviews are prohibitively time consuming. Thus, self-administered screening questionnaires such as the MDQ have potential clinical utility insofar as they can improve the efficiency of the diagnostic evaluation. There are no special questions on the MDQ that allow it to detect a history of mania or hypomania that otherwise would go undetected during a clinical evaluation. For that matter, there are no special questions on the SCID, and the content and wording of the SCID's questions are recognizable to nonresearcher clinicians. However, completing the entire SCID can take several hours, and this simply is not feasible for most clinicians.

Received April 23, 2003; accepted Aug. 4, 2003. From the Department of Psychiatry and Human Behavior, Brown University School of Medicine, Rhode Island Hospital, Providence, R.I.

This research was supported, in part, by grants MH48732 and MH56404 from the National Institute of Mental Health, Bethesda, Md.

Corresponding author and reprints: Mark Zimmerman, M.D., Bayside Medical Center, 235 Plain Street, Providence, RI 02905 (e-mail: mzimmerman@lifespan.org).

Figure 1. Table of Association Between a Diagnostic Gold Standard and Screening Test

Gold Standard Diagnosis

Screening Test	Present	Absent	Total
Positive	a	b	a + b
Negative	c	d	c + d
Total	a + c	b + d	a + b + c + d

Sensitivity = a/(a + c)
Specificity = d/(b + d)
Positive Predictive Value = a/(a + b)
Negative Predictive Value = d/(c + d)
Overall Correct = (a + d)/(a + b + c + d)

$$\text{Kappa} = \frac{\text{(observed agreement} - \text{chance agreement)}}{(1 - \text{chance agreement})}$$

REVIEW OF THE STATISTICS OF DIAGNOSTIC TESTING

There are several excellent articles describing the descriptive statistics of test performance.[6–10] Despite these, in several studies of the performance of self-administered screening tests, incorrect definitions and miscalculations of these statistics were found[11]; therefore, we present a brief overview of this area.

In studies of the performance of self-report screening questionnaires, 6 statistics are most commonly computed: sensitivity, specificity, positive predictive value, negative predictive value, overall level of agreement, and chance corrected level of agreement. These statistics are easy to compute when a table of association between the test and the gold standard is presented, as in Figure 1.

The *sensitivity* (true positive rate) of a screening test refers to how well the test identifies individuals with the illness. When computing sensitivity, the numerator is the number of ill persons who are correctly identified as ill by the test, and the denominator is the total number of ill persons [$a/(a + c)$].

Specificity (true negative rate) refers to how well the screening test identifies individuals without the illness. When computing specificity, the numerator is the number of persons without the illness who are correctly identified by the test as not having the illness, and the denominator is the total number of persons without the illness [$d/(b + d)$].

The predictive values of a test are generally more clinically meaningful than a test's sensitivity and specificity. The predictive value statistics indicate the probability that an individual is ill or not ill, given that the screening test identifies them as ill or not ill. Accordingly, *positive predictive value* refers to the probability that a person who is identified as ill by the test actually has the illness. When computing positive predictive value, the numerator is the same as it is in computing sensitivity (i.e., the number of ill persons who are correctly identified as ill by the test); however, the denominator is now the total number of persons that the test identifies as ill [$a/(a + b)$].

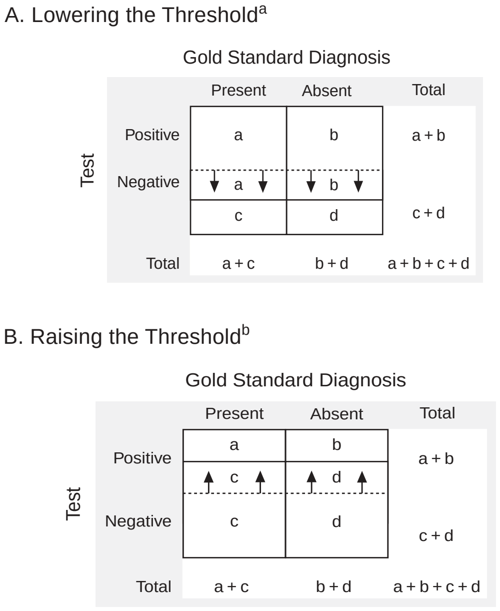
Negative predictive value refers to the probability that a person who is identified as not ill by the test actually is not ill. When computing negative predictive value the numerator is the same as it is in computing specificity (i.e., the number of not ill persons who are correctly identified as not ill by the test), though the denominator is the total number of persons that the test identifies as not ill [$d/(c + d)$].

The *absolute* or *overall level of agreement* (sometimes referred to as the *hit rate*) refers to the total number of patients who are correctly classified by the test [$(a + d)/(a + b + c + d)$]. *Kappa* represents the level of agreement beyond that accounted for by chance. Other statistics have also been used to correct for chance agreement,[12] but kappa is the most frequently used.

Sensitivity, specificity, and positive and negative predictive values are not invariant properties of a test—they can be manipulated by changing the cutoff point used to distinguish cases from noncases and are influenced by disease prevalence. Four axioms characterize the relationships between these statistics.

1. Lowering a test's cutoff score to identify cases increases the test's sensitivity and decreases its specificity. As illustrated in Figure 2A, when the cutoff score (above which individuals are designated cases and below which they are classified as noncases) is lowered, some persons previously in cells *c* and *d* of the 2 × 2 matrix will be redistributed into cells *a* and *b*, respectively. Sensitivity [$a/(a + c)$] increases because cell *a* increases while the sum of $a + c$ remains the same (because the true disease prevalence is unaffected by the test threshold). Similarly, specificity [$d/(b + d)$] decreases when the threshold is lowered because *d* decreases and $b + d$ remains the same.
2. Conversely, raising the test threshold to identify cases decreases the test's sensitivity and increases its specificity. Following the same logic as above, when the threshold increases, sensitivity decreases because the size of cell *a* decreases, and specificity increases because *d* increases. This is illustrated in Figure 2B.
3. A test's positive predictive value is higher in samples where disease prevalence is greater. This postulate assumes that test sensitivity and specificity are fixed across samples and independent of prevalence. Consider 2 studies with samples of equal size but different illness prevalence rates.

Figure 2. Illustration of the Impact of Lowering and Raising the Screening Test Cutoff Score on the Cell Sizes of the Table of Association Between a Diagnostic Gold Standard and Screening Test



A. Lowering the Threshold[a]

B. Raising the Threshold[b]

[a]When the threshold is lowered (from the dashed line to the solid line), then cells *a* and *b* increase, and cells *c* and *d* are reduced a corresponding amount.

[b]When the threshold is raised (from the dashed line to the solid line), then cells *a* and *b* decrease, and cells *c* and *d* are increased a corresponding amount.

When test sensitivity [*a*/(*a* + *c*)] is the same, then in the sample with the higher prevalence (*a* + *c*), both cells *a* and *c* must be greater. Likewise, when specificity [*d*/(*b* + *d*)] remains the same, then when wellness (*b* + *d*) decreases, both cells *b* and *d* must be lower. Positive predictive value [(*a*/(*a* + *b*)] is greater in the sample with a higher prevalence because *b* is smaller.

4. At constant sensitivity and specificity, a test's negative predictive value is higher in samples where disease prevalence is lower. When sensitivity [*a*/(*a* + *c*)] is the same in 2 samples, then in the sample with lower disease prevalence (*a* + *c*), both *a* and *c* must be lower. When specificity [*d*/(*b* + *d*)] is the same in 2 samples, then when more individuals are not ill (*b* + *d*), both *b* and *d* increase. Thus, negative predictive value [*d*/(*c* + *d*)] is higher in samples with lower rates of disorder (i.e., higher rates of nondisorder) because *c* is smaller.

Depending on the instrument's purpose, cutoff scores might be selected to optimize the sensitivity or specificity of the scale.[13,14] To optimize sensitivity, the threshold to identify cases will be lowered, whereas to optimize specificity, the threshold will be raised. Investigators can also alter disorder prevalence when examining the performance of a screening measure. Disorder prevalence is not usually thought to be under the control of the investigator; however, prevalence is a function of the broadness of the definition of disorder. For example, studies of the diagnostic performance of depression questionnaires usually focus on identifying cases of major depressive disorder (MDD). Some studies, though, have used a broader definition of depression and included dysthymic disorder, minor depression, and adjustment disorder with MDD.[15,16] This issue will be discussed further in our review of the studies of the MDQ.

From a clinical perspective, it is most important that a screening scale have good sensitivity (thus the threshold should be set low) even if that means that false positives are increased because specificity is lower. False positives are less of a problem for a psychiatric screening questionnaire because their major cost is the time a clinician takes to determine that the disorder is not present. Presumably, this is time the clinician would have nonetheless spent for the same purpose. Also, the measure should have high negative predictive value. With high negative predictive value, the clinician can be confident that when the test indicates that the disorder is not present there is little need to inquire about that disorder's symptoms. A test with high negative predictive value therefore can help focus the initial evaluation.

DESCRIPTION OF THE MDQ

The MDQ was printed in the initial article describing its development and validation.[2] The scale screens for a lifetime history of mania or hypomania with 13 yes/no symptom questions reflecting the DSM-IV inclusion criteria. The symptom questions are followed by a single yes/no question about whether symptoms clustered during the same period of time. The respondent is instructed to answer this question only if more than 1 symptom was checked off. The final question evaluates the level of impairment resulting from the symptoms. This item is rated on a 4-point scale (no problem, minor problem, moderate problem, serious problem). On the basis of the results of the initial validation study, a positive case requires the presence of 7 or more symptom items that cluster within the same time period and cause either moderate or serious problems.

STUDIES OF THE MDQ

The first study of the MDQ was conducted in psychiatric clinics that specialize in the treatment of mood disorders.[2] One hundred ninety-eight patients completed the MDQ and were interviewed with the mood disorders module of the SCID. The majority of the patients in the sample were diagnosed with bipolar disorder according to the SCID (N = 109, 55.1%), and the majority of the patients with bipolar disorder had bipolar I disorder (N = 70, 64.2%). The authors examined the sensitivity and specificity of the MDQ at every cutoff and found that a cutoff of 7 provided the best balance between sensitivity and specificity. Figure 1 of their article presented the sensitivity and specificity of the MDQ at every threshold score from 0 through 13. Apparently there is a mistake in this figure: at a score of 0, the sensitivity should have been close to 100% and the specificity close to 0%. Instead, sensitivity is shown at 80% and specificity at 68%. Following from the previous discussion on the impact of lowering the threshold on sensitivity and specificity, reducing the cutoff to 0 should result in almost all patients being considered as cases on the MDQ because almost everyone would fall into cells *a* and *b* of the 2 × 2 table of association (Figure 1). (Because MDQ caseness depends on the number of symptom items checked as well as symptom clustering in the same time period and the presence of at least moderate distress, we cannot assert that every patient with a symptom score of 0 would fall into cells *a* and *b*. In fact, it is unclear how a patient would answer the question about impairment if they did not check off any symptom items.) Specificity should be close to 0% because few patients should be in cell *d*.

A limitation of the initial study of the MDQ was that it was conducted with patients already being treated in specialized mood disorders clinics at academic medical centers that had a particular expertise in treating bipolar disorder. It could be expected that patients in ongoing treatment in these clinics had received psychoeducation and were more aware of their diagnoses than are patients usually treated in the community. Also, the prestige of these clinics may have resulted in more prototypical and refractory cases being treated. Because of these factors we would predict that the performance of the MDQ would be significantly poorer in a replication study conducted in a different type of setting. As summarized later, this was in fact found to be true in a second study of community participants.[3] Moreover, almost two thirds of the patients with bipolar disorder were diagnosed with bipolar I disorder. In a critique of the study, Mago[17] suggested that it is easier to diagnose bipolar I disorder than milder bipolar spectrum disorders and predicted that the sensitivity of the MDQ would be lower for the more subtle bipolar spectrum disorders. In his reply to Mago, Hirschfeld[18] did not indicate whether the diagnostic performance of the MDQ differed for bipolar I disorder than the other bipolar disorders. Clinically, bipolar I disorder is more readily recognized and diagnosed than are more subtle bipolar disorders because a diagnosis of mania requires hospitalization, psychosis, or gross functional impairment. Thus, the principal potential clinical utility of the MDQ lies with its ability to detect these milder bipolar syndromes.

The second study of the diagnostic performance of the MDQ was conducted in a subset of subjects from the epidemiologic study. Six hundred ninety-five subjects completed the MDQ and were interviewed by phone with the mood disorders module of the SCID. Subjects were selected for this study based on their MDQ score to ensure adequate numbers of high scorers on the MDQ. Thus, this subsample was, by design, not representative of the larger epidemiologic study in terms of the distribution of MDQ scores. More than one tenth of this sample was diagnosed with bipolar disorder according to the SCID (11.2%, N = 78), the vast majority of whom had bipolar I disorder (N = 70). The data from the subsample were weighted to represent the demographic characteristics of the larger sample. The authors reported that the sensitivity and specificity was examined for all cutoff points on the MDQ, though data were only reported for a cutoff of 7. The weighted sensitivity and specificity of the MDQ were 28.1% and 97.2%, respectively. The authors concluded that the sensitivity was “less than ideal” and attributed this to the modest test-retest reliability of the SCID in the general population. It is true that less-than-perfect reliability puts a ceiling on the agreement between a self-report scale and the diagnostic interview; however, there is no evidence that the test-retest reliability of diagnosing bipolar disorder is different in patient and nonpatient samples.[19] Therefore, it is unclear why, on the basis of diagnostic reliability, the sensitivity would be expected to be so much lower in this study than the patient study. Our hypothesis, as noted previously, is that the sensitivity of the MDQ in the original patient sample was artificially inflated because the subjects were recruited from specialized mood disorder clinics and were more prototypical, severely ill, and educated about their diagnosis.

Despite the low sensitivity in the second study,[3] the MDQ was subsequently used as a case-finding instrument in the large epidemiologic study of more than 85,000 subjects.[1] The problem with using the MDQ as a case-finding measure is that the scale lacks sufficient diagnostic precision. To illustrate, we generated a hypothetical 2 × 2 table of association for 1000 subjects (Figure 3). In this table we assumed that the prevalence of bipolar disorder according to the SCID was 3.7% (the weighted prevalence rate based on the MDQ). (If we had used a more conventional prevalence estimate of 1.5%, then the MDQ would perform more poorly than in our example.) A prevalence of 3.7% means that 37 subjects have bipolar disorder according to the SCID, and 963 would not be diagnosed with bipolar disorder. With an MDQ sensitivity of 28.1%, 10 of 37 bipolar subjects according to the SCID would screen positive on the MDQ. With a specificity of 97.2%, 934 of the 963 subjects who were not bipolar on the SCID would screen negative on the MDQ, and 29 subjects would screen positive on the MDQ. Figure 3 shows that 66 subjects would be identi-

Figure 3. Hypothetical Table of Association Between the MDQ and SCID in 1000 Community Subjects Based on an MDQ Sensitivity of 28.1% and Specificity of 97.2%

		SCID Bipolar Disorder Diagnosis		
		Present	Absent	Total
MDQ Caseness	Positive	10	29	39
	Negative	27	934	961
	Total	37	963	1000

Sensitivity = a/(a + c) = 10/37 = 27.0%

Specificity = d/(b + d) = 934/963 = 97.0%

Positive Predictive Value = a/(a + b) = 10/39 = 25.6%

Negative Predictive Value = d/(c + d) = 934/961 = 97.2%

Overall Correct = (a + d)/(a + b + c + d) = 944/1000 = 94.4%

$$\text{Kappa} = \frac{(\text{observed agreement} - \text{chance agreement})}{(1 - \text{chance agreement})} = \frac{(0.944 - 0.927)}{(1 - 0.927)} = 0.23$$

Abbreviations: MDQ = Mood Disorders Questionnaire, SCID = Structured Clinical Interview for DSM-IV.

fied as bipolar by at least one of the instruments; however, only 10 of these 66 would be identified as bipolar by both the SCID and MDQ. Thus, while the prevalence of bipolar disorder is nearly the same according to the 2 measures, for the most part, they identify different individuals. Consistent with this, the kappa coefficient of agreement between the MDQ and SCID is 0.23, indicating poor agreement.

The above analysis raises questions about the adequacy of the MDQ as a case-finding instrument in the community. There is little overlap in the subjects who are identified as bipolar by the MDQ and SCID, and the positive predictive value of the MDQ is low (25.6%; 10/39). Most MDQ cases of bipolar disorder would not be diagnosed as such on the SCID.

We believe that the limited diagnostic validity of the MDQ is responsible for some of the results of the epidemiologic study. The prevalence of MDQ bipolar disorder was disproportionately high in the 18- to 24-year-old age cohort, more than twice as high as in the 25- to 34-year-old age cohort. We suspect that the mood and behavioral instability of cluster B personality pathology and substance use disorders, which are associated with young age,[20,21] resulted in a high rate of false positives in the youngest age cohort.

Hirschfeld and colleagues[1] also found that 80% of MDQ cases had not been previously diagnosed with bipolar disorder by a doctor and suggested that bipolar disorder is underrecognized. They found that 31% of MDQ cases had been previously diagnosed with unipolar depression, and discussed the risks of prescribing antidepressants to depressed bipolar patients. However, the low positive predictive value of the MDQ (approximately 25%) suggests that the problem may not be with clinician underrecognition but may be with MDQ false positives.

In discussing the limitations of the epidemiologic study, Hirschfeld et al.[1] duly noted that the MDQ is a screening scale, not a diagnostic measure. However, to support their use of the MDQ as a case-finding instrument, they concluded that the scale had "generally good sensitivity and specificity with regard to research diagnostic interviews obtained from trained interviewers in both clinical and, in the current report, nonclinical samples."[1(p58)] We respectfully disagree. We believe that the sensitivity of the MDQ is too low for routine use as a screening measure in clinical samples (73%) and even less satisfactory in nonclinical samples (28%). Hirschfeld et al. further suggested that the MDQ can be used as a case-finding measure in the nonclinical, epidemiologic study because the prevalence of bipolar disorder was nonsignificantly different according to the MDQ and SCID. However, as demonstrated in Figure 3, disorder prevalence on a self-report scale and interview can be equivalent yet identify different individuals as cases. Disorder prevalence on a self-report screening scale is simply a function of where the cutoff is set to identify cases. A good case-finding instrument in the general population not only yields correct prevalence estimates, but also must demonstrate a high level of agreement with a gold standard instrument. That this is possible was demonstrated by one of us several years ago with a self-report depression scale designed to "diagnose" major depressive disorder.[22]

CONCLUSIONS

Bipolar disorder is a serious illness that has an early age at onset[23] and is subsequently associated with chronic morbidity,[24,25] frequent recurrences of mood syndromes,[26] high utilization of the mental health system,[27] impairment in occupational functioning[28] and interpersonal and family relationships,[29,30] and suicidality.[31] In light of the multiple and severe public health issues presented by patients with bipolar disorder, improving its recognition seems a worthwhile effort. While we applaud the efforts of Hirschfeld and colleagues to develop a questionnaire to improve the detection of bipolar disorder, the ability of the MDQ to perform its intended task is modest, at best.

We conclude from our examination of the diagnostic properties of the MDQ that the scale is not sufficiently accurate to be used as either a case-finding measure in community studies or a screening scale in clinical practice. Additional study of the measure, and perhaps a modified version of the scale, is warranted. An examination of the MDQ's diagnostic properties in clinical sites that are not specialized centers for the treatment of mood disorders would provide important data on the scale's performance in routine clinical practice. Such a study could examine whether the measure performs similarly in identifying bipolar I and bipolar II disorder and whether the cutoffs should differ in identifying hypomania and mania. Further study of the MDQ in nonpatient samples should examine the possible causes of the scale's low sensitivity. The low positive predictive value of the MDQ renders it inappropriate for use as a case-finding measure for bipolar disorder in the community because the majority of MDQ positives are not diagnosed with bipolar disorder according to a standardized diagnostic interview.

REFERENCES

1. Hirschfeld R, Calabrese J, Weissman M, et al. Screening for bipolar disorder in the community. J Clin Psychiatry 2003;64:53–59
2. Hirschfeld R, Williams J, Spitzer R, et al. Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. Am J Psychiatry 2000;157:1873–1875
3. Hirschfeld R, Holzer C, Calabrese J, et al. Validity of the Mood Disorder Questionnaire: a general population study. Am J Psychiatry 2003;160: 178–180
4. Griner P, Glaser R. Misuse of laboratory tests and diagnostic procedures. N Engl J Med 1966;307:1336–1339
5. First MB, Spitzer RL, Gibbon M, et al. Structured Clinical Interview for DSM-IV Axis I Disorders, Patient Edition (SCID-I/P, version 2.0). New York, NY: Biometrics Research, New York State Psychiatric Institute; 1995
6. Baldessarini RJ, Finklestein S, Arana GW. The predictive power of diagnostic tests and the effect of prevalence of illness. Arch Gen Psychiatry 1983;40:569–573
7. Fletcher R, Fletcher S, Wagner E. Clinical Epidemiology. Baltimore, Md: Williams & Wilkins; 1988
8. Griner PF, Mayewski RJ, Mushlin AI, et al. Selection and interpretation of diagnostic tests and procedures: principles and applications. Ann Int Med 1981;94:553–600
9. Mausner J, Kramer S. Epidemiology: An Introductory Text. Philadelphia, Pa: W. B. Saunders; 1985
10. Sackett D. A primer on the precision and accuracy of the clinical examination. JAMA 1992;267:2638–2644
11. Kessel JB, Zimmerman M. Reporting errors in studies of the diagnostic performance of self-administered questionnaires: extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. Psychol Assess 1993;5:395–399
12. Spitznagel E, Helzer J. A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 1985;42:725–728
13. Mossman D, Somoza E. Maximizing diagnostic information from the dexamethasone suppression test. Arch Gen Psychiatry 1989;46:653–660
14. Hsiao JK, Bartko JJ, Potter WZ. Diagnosing diagnoses: receiver operating characteristic methods and psychiatry. Arch Gen Psychiatry 1989;46:664–667
15. Rounsaville BJ, Weissman MM, Rosenberger PH, et al. Detecting depressive disorders in drug abusers: a comparison of screening instruments. J Affect Disord 1979;1:255–267
16. Willenbring ML. Measurement of depression in alcoholics. J Stud Alcohol 1986;47:367–372
17. Mago R. Bipolar disorder questionnaire [letter]. Am J Psychiatry 2001;158:1743
18. Hirschfeld R. Bipolar disorder questionnaire [letter]. Am J Psychiatry 2001;158:1743–1744
19. Simpson S, McMahon F, McInnis M, et al. Diagnostic reliability of bipolar II disorder. Arch Gen Psychiatry 2002;59:736–740
20. Mattia J, Zimmerman M. Epidemiology of personality disorders. In: Livesley J, ed. Handbook of Personality Disorders. New York, NY: Guilford Press; 2001:107–123
21. Kessler RC, McGonagle KA, Zhao S, et al. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Arch Gen Psychiatry 1994;51:8–19
22. Zimmerman M, Coryell W. Screening for major depressive disorder in the community: a comparison of measures. Psychol Assess 1994;6:71–74
23. Burke K, Burke J Jr, Regier D, et al. Age at onset of selected mental disorders in five community populations. Arch Gen Psychiatry 1990;47:511–518
24. Judd L, Akiskal H, Schettler P, et al. The long-term natural history of the weekly symptomatic status of bipolar I disorder. Arch Gen Psychiatry 2002;59:530–537
25. Judd L, Akiskal H, Schettler P, et al. A prospective investigation of the natural history of the long-term weekly symptomatic status of bipolar II disorder. Arch Gen Psychiatry 2003;60:261–269
26. Keller M, Lavori P, Coryell W, et al. Bipolar I: a five-year prospective follow-up. J Nerv Ment Dis 1993;181:238–245
27. Bourdon KA, Rae DS, Narrow WE, et al. National prevalence and treatment of mental and addictive disorders. In: Manderscheid RW, Sonnenschein MA, eds. Mental Health, United States. DHHS Publication No. (SMA) 94-3000. Washington, DC: U.S. Government Printing Office; 1994:22–51
28. Tohen M, Waternaux C, Tsuang M. Outcome in mania: a 4-year prospective follow-up of 75 patients utilizing survival analysis. Arch Gen Psychiatry 1990:47:1106–1111
29. Coryell W, Scheftner W, Keller M, et al. The enduring psychosocial consequences of mania and depression. Am J Psychiatry 1993;150: 720–727
30. Chakrabarti S, Kulhara P, Verma S. Extent and determinants of burden among families of patients with affective disorders. Acta Psychiatr Scand 1992;86:247–252
31. Goodwin F, Jamison K. Manic-Depressive Illness. New York, NY: Oxford University Press; 1990

See Hirschfeld and colleagues' response to this commentary on page 721.