Screening for Bipolar Disorder: Confusion between Case-Finding and Screening

Mark Zimmerman

Department of Psychiatry and Human Behavior, Brown Medical School, and Department of Psychiatry, Rhode Island Hospital, Providence, R.I., USA

Introduction

Bipolar disorder is a serious illness resulting in significant psychosocial morbidity and excess mortality [1]. The prevalence of bipolar disorder is relatively high, especially when it is defined to include milder variants such as bipolar II disorder and bipolar disorder not otherwise specified [2].

Despite its clinical and public health significance, studies of psychiatric and primary care patients have found that bipolar disorder is often missed [3–5]. The underdiagnosis of bipolar disorder in depressed patients has potential treatment and clinical implications such as the underprescription of mood-stabilizing medications, an increased risk of rapid cycling due to possible overprescription of antidepressant medications and increased costs of care [3, 6–8].

Several causes of bipolar disorder underdiagnosis have been postulated. Patients with bipolar disorder are more likely to experience symptoms of depression and anxiety, and present for treatment of these symptoms rather than symptoms of mania or hypomania [9, 10]. Because patients with bipolar disorder usually do not present in the manic or hypomanic phase of the illness, to correctly diagnose bipolar disorder, the presence of a prior history of hypomanic or manic episodes must be elicited. Recommendations for improving the detection of bipolar disorder include careful clinical evaluations, inquiring about a history of mania and hypomania, and the use of screening questionnaires [11–13].

Many scales have been developed to assess manic/hypomanic symptoms. Some of these instruments are cross-sectional measures of symptom severity, and, prior to DSM-5, would not have been appropriate as diagnostic aids in currently depressed patients. The reason for the caveat regarding DSM-5 is because DSM-5 includes a mixed feature specifier for depressed patients which these scales might be useful in identifying. In the past few years, 4 self-report questionnaires have been developed to screen for bipolar disorder [14–17]. That is, they assess a lifetime history of manic/hypomanic symptoms. The purpose of the present article is twofold. First, I briefly describe the structure and content of these 4 scales and summarize the initial studies establishing the cutoff score derived by the scale developers that was recommended to screen for bipolar disorder. Second, I discuss the distinction between case-finding and screening, how this distinction is reflected in the derivation of a measure’s cutoff score, and examine how this applies to the 4 bipolar disorder screening instruments.
Description of Four Screening Scales for Bipolar Disorder

The Hypomanic Checklist

The goal of the Hypomanic Checklist (HCL-32) was to improve the recognition of hypomanic features in depressed patients and thereby enhance recognition of bipolar II disorder and other bipolar spectrum disorders [16]. The introduction to the HCL-32 notes: ‘At different times in their life everyone experiences changes or swings in energy, activity and mood (“highs and lows” or “ups and downs”). The aim of this questionnaire is to assess the characteristics of the “high” periods.’ The respondent is then instructed to think of a ‘period when you were in a “high” state’ and answer 32 yes-no questions about their mood and behavior during that time. Each no response is scored 0 and each yes response is scored 1, and the total HCL-32 score is obtained by summing the number of yes responses. Thus, the range of scores spans from 0 to 32. There are some additional items to the scale assessing other clinically important variables such as functional impairment and symptom duration, but these do not contribute to the total score and do not contribute to determining whether the respondent screens positive. The scale is reprinted in the published article.

The initial study of the HCL-32 was of 446 Italian and Swedish patients (all but 7 were outpatients) with major depressive disorder or bipolar disorder. The authors recommended a cutoff score of 14 because it yielded the best combination of sensitivity (80%) and specificity (51%) in distinguishing bipolar disorder from major depressive disorder. The authors did not present the sensitivity and specificity values for other cutoff scores.

Bipolar Spectrum Diagnostic Scale

The authors of the Bipolar Spectrum Diagnostic Scale (BSDS) had conducted a previous study of the Mood Disorders Questionnaire and concluded that the scale had good sensitivity for bipolar I disorder (70%) but a less impressive sensitivity for bipolar II disorder and bipolar disorder not otherwise specified [18]. The goal of the BSDS was to improve the recognition of milder, subthreshold, forms of the bipolar spectrum and thus be effective across the entire bipolar spectrum severity range [19]. The scale consists of two parts. The first part is a paragraph containing 19 statements describing many of the symptoms of bipolar disorder, with each sentence followed by an underlined space for respondents to place a checkmark if they believe that the statement applies to them. In scoring the scale, each item checked is counted as 1 point. The second part of the BSDS is a single multiple-choice question asking respondents how well the paragraph describes them [very well, or almost perfectly (6 points); fairly well (4 points); to some degree, but not in most respects (2 points); not really at all (0 points)]. Thus, the range of scores spans from 0 to 25. The instructions of the BSDS ask patients to read through the entire paragraph before checking off which symptom statements apply to them. After reading the paragraph, patients are asked to complete the multiple-choice question, and then to go back and put a check mark after each statement that ‘definitely describes you’. A copy of the BSDS is reprinted in the article by Ghaemi et al. [19].

The initial study of the BSDS was of 71 patients with major depressive disorder or bipolar disorder. A priori the authors suggested that a cutoff of 12 represented a positive screen. At this cutoff the sensitivity of the BSDS was 76%, and its specificity was 85%. They examined the performance of the scale at other cutoff points and recommended a cutoff score of 13 because it yielded the best combination of sensitivity (75%) and specificity (93%) in distinguishing bipolar disorder from major depressive disorder. The authors presented the sensitivity and specificity values for cutoff scores of 10 through 14 in a figure.

Mood Swings Questionnaire/Survey

In the initial publication this scale was referred to as the Mood Swings Survey [20], but more recently it has been called the Mood Swings Questionnaire (MSQ), and I will therefore use this name [21, 22].

The goal of the MSQ was to improve the recognition of bipolar II disorder in depressed patients and to better distinguish bipolar I and bipolar II disorder. As described in early publications [21], the MSQ begins with a 2-part screening question: ‘Do you ever have mood swings, and as part of such swings, have times when (i) your mood is higher than your usual sense of happiness (“a high”) and (ii) you feel quite “wired”, “energized”, “elevated”, “expansive”, and possibly “irritable”? If the respondent answers no to this question, the remainder of the scale is not completed. (In more recent descriptions, including the version available on the Black Dog Institute website, this complex question has been simplified. However, the approach remains the same insofar as the subsequent questions are not completed if the screening question is answered in the negative.) This question is followed by 46 questions rated on a 3-point scale [no more than usual (scored 0); somewhat more than usual (scored 1); much more than usual (scored 2)]. A 27-item version of the scale has also been developed. The range of scores spans...
from 0 to 92 (0 to 54 for the shorter version). The briefer 27-item scale is reprinted on the Black Dog Institute website.

The initial study of the MSQ-46 was of 157 depressed outpatients. The authors recommended a cutoff score of 36 as optimal in distinguishing bipolar disorder and major depressive disorder. At this cutoff the scale achieved a sensitivity of 84% and a specificity of 93%. For the shortened 27-item version, the optimal cutoff of 22 resulted in a sensitivity of 81% and a specificity of 98%. The authors did not present the sensitivity and specificity values for other cutoff scores.

**Mood Disorders Questionnaire**

The Mood Disorders Questionnaire screens for a lifetime history of mania or hypomania with 13 yes/no symptom questions reflecting the DSM-IV inclusion criteria [15]. Total scores are calculated by summing the 13 items that are answered yes. The symptom questions are followed by a single yes/no question about whether the 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).

The initial study of the MDQ was of 198 outpatients treated in clinics specializing in the treatment of mood disorders, especially bipolar disorder. The authors recommended a cutoff score of 7 symptoms that co-occurred and caused moderate or serious problems because it yielded the best combination of sensitivity (73%) and specificity (90%) in distinguishing bipolar disorder from major depressive disorder. The authors presented a graph of the sensitivity and specificity of the scale for all symptom cutoff scores.

**Screening or Case-Finding?**

The threshold chosen to identify cases on a symptom questionnaire with a continuous score distribution is based on the intended use of the scale. If the goal is to identify a relatively homogeneous group of individuals who are highly likely to have the index disorder, a high threshold will be chosen in order to increase the specificity of the scale and reduce the number of false-positive results in the group scoring above the threshold. To achieve this, high sensitivity is sacrificed. On the other hand, if the purpose is to use the scale as screening measure, as purportedly each of these tests was intended to be, then a broad net needs to be cast and the threshold set low in order to enhance the sensitivity of the measure. It is important for a screening scale to have high sensitivity because the more time-intensive follow-up diagnostic inquiry will presumably only occur in patients who are positive on the initial screen. Clinicians who rely on the screening scale as the first stage of a 2-stage diagnostic assessment, with the second stage being the more in-depth diagnostic interview, will miss the diagnosis in patients who are false-negative on the instrument.

In developing each of these scales, the 4 research groups recommended a cutoff score that provided the best balance between sensitivity and specificity thereby optimizing agreement between the test and the diagnostic standard. While there is nothing inherently wrong with this method of selecting a cutoff score, it is at odds with the stated intended use of the scale as a screening instrument. In the initial validation study of 3 of the 4 scales, the sensitivity of the measure at the recommended cutoff score was lower than its specificity.

None of the scale developers discussed how their test should be used in real-world clinical practice. Consistently with the concept of screening, the scales should be administered prior to the clinical interview, with the clinician following up positive screens. Did the authors envision an alternative manner of using the scale? Did the authors expect clinicians to inquire about a history of manic and hypomanic episodes for patients who screened negative on the scale? If not, then scale sensitivity is of critical importance. If so, then what is the purpose of the scale? Do the authors of these scales believe that clinicians do not know what questions to ask to assess prior manic/hypomanic episodes and therefore these scales are needed because the scales ask the right questions that clinicians do not? Presumably the scale developers expected clinicians to follow up positive screens with a diagnostic interview because they did not suggest that these self-administered scales are definitive diagnostic tests. Moreover, from a feasibility of clinical use perspective, none of the scale developers attempted to develop as brief a measure as possible, which would be desirable in a screening instrument that could be used alongside measures that screen for other disorders that are important to recognize in depressed patients. A screening measure does not need to be comprehensive and cover all of the criteria of mania/hypomania, but instead needs to include a sufficient number of items to achieve a high sensitivity and acceptable specificity.
None of the scale developers discussed the possibility of false-positive diagnoses associated with false-positive screening test results which would occur because a clinician was overly influenced by the results of the screening test. False-positive cases are at risk of overtreatment with mood stabilizers. For other medical disorders, screening recommendations typically include a discussion of both potential benefits and harms. There are several examples in the research literature in which the results of bipolar disorder screening tests were misinterpreted as suggestive of actual caseness rather than screen positives [23]. If researchers (as well as the reviewers of these articles published in peer review journals) misinterpreted the results of these tests, are clinicians likely to do so as well?

Recommendations to use screening scales for bipolar disorder in clinical practice are common [24–27]. While the underdiagnosis of bipolar disorder is a problem in clinical practice, a balanced discussion of the potential risks and benefits of screening has been lacking. It is important for the developers of screening scales, and investigators of the performance of these measures, to discuss their rationale for selecting cutoff scores that maximize agreement with a diagnostic standard rather than selecting a cutoff score that favors sensitivity over specificity.

Disclosure Statement

The author has no conflict of interest to declare.