Speaking a More Consistent Language When Discussing Severe Depression: A Calibration Study of 3 Self-Report Measures of Depressive Symptoms

Mark Zimmerman, MD; Jennifer H. Martinez, BA; Michael Friedman, MD; Daniela A. Boerescu, MD; Naureen Attiullah, MD; and Cristina Toba, MD

ABSTRACT

Objective: We recently found marked disparities between 3 self-report scales that assess the *DSM-IV* criteria for major depressive disorder in the percentage of depressed outpatients considered to have severe depression. The goal of the present report from the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project was to calibrate the measures against a clinician-rated criterion standard and to establish a cutoff point on each scale that identifies a similar prevalence of severe depression and increases the level of agreement between the scales in identifying severe depression.

Method: 353 depressed outpatients (*DSM-IV*) completed the Clinically Useful Depression Outcome Scale, Quick Inventory of Depressive Symptomatology, and Patient Health Questionnaire from June 2010 to January 2013. The patients were also rated on the 17-item Hamilton Depression Rating Scale (HDRS). The goal of the analyses was to identify the cutoff point on each of the self-report scales that would identify a prevalence of severe depression similar to that identified by the HDRS (defined as a score of 25 and above).

Results: On the basis of the scale developers' recommended cutoffs, the prevalence of severe depression varied greatly (range, 15.3%–67.4%), and the level of agreement between the pairs of scales was low. After calibration, the self-report scales identified a similar percentage of patients as severely depressed (range, 22.2%–26.5%), and the level of agreement between the scales in identifying severe depression increased.

Discussion: If clinicians are to follow treatment guidelines' recommendations to base initial treatment selection, in part, on depression severity, then it is important to have a consistent method of determining depression severity. The present calibration study of 3 self-report depression questionnaires identified cutoff scores that resulted in similar prevalence rates of severe depression and increased the level of agreement between the scales.

J Clin Psychiatry 2014;75(2):141–146

Submitted: *March 7, 2013; accepted June 24, 2013.*
Online ahead of print: *December 10, 2013 (doi:10.4088/JCP.13m08458).*

Corresponding author: *Mark Zimmerman, MD, 146 West River St, Providence, RI 02904 (mzimmerman@lifespan.org).*

When treating patients with major depressive disorder (MDD) in clinical practice, it is important to measure severity because depression severity predicts treatment outcome and should be considered in treatment selection. Greater symptom severity is associated with a higher response to antidepressant medication, a lower response to placebo, and, thus, a greater separation between active drug and placebo response.[1–3] In severely depressed patients, response to psychotherapy has been found to be inferior to medication response,[2] although a recent meta-analysis[4] of psychotherapy studies found that greater symptom severity did not predict poorer response in controlled studies examining the moderating effect of severity. Certain medications or classes of medication have been hypothesized to be more effective than others for severe depression, though this differential effectiveness has not received consistent empirical support.[5–9]

The recently revised American Psychiatric Association (APA) guidelines[6] for the treatment of MDD indicate that it is important to consider symptom severity in initial treatment selection. Specifically, the guidelines recommend both psychotherapy and pharmacotherapy as monotherapies for mildly and moderately severe depression and pharmacotherapy with or without psychotherapy for severely depressed patients. Guidelines from other countries have also recommended pharmacotherapy as the first treatment option for severely depressed patients and either pharmacotherapy or psychotherapy for mildly and moderately depressed patients.[10,11]

In addition to making recommendations for treatment approach based on severity, the APA's revised treatment guidelines for MDD advocate the use of standardized, quantitative measures to evaluate treatment outcome. Reliable and valid self-report questionnaires may be preferable to clinician-rated scales such as the Hamilton Depression Rating Scale (HDRS)[12] or the Montgomery-Asberg Depression Rating Scale[13] because they are inexpensive in terms of professional time needed for administration.

Because of the significance accorded severity by treatment guidelines, our clinical research group compared different scales based on their allocation of patients to severity groups.[14] We found marked disparities between the Clinically Useful Depression Outcome Scale (CUDOS),[15] Quick Inventory of Depressive Symptomatology (QIDS),[16] and Patient Health Questionnaire (PHQ-9)[17] in the percentage of depressed outpatients considered to have severe depression.[14] In the face of such disparity, we thought it would be difficult to convince clinicians to use such measures to guide treatment selection, and it would be difficult to evaluate how well practice standards are being followed in the treatment of severe depression. Because of the more than 2-fold difference between the scales in the percentage of patients considered to have severe depression, we

Clinical Points

- Treatment guidelines for depression suggest that severity should be taken into account when initiating treatment.
- There is a marked disparity between self-report scales in the classification of depressed outpatients into severity groups because the authors of the scales used different methods to derive cutoff points to identify severe depression.
- Calibrating the self-report scales against the Hamilton Depression Rating Scale identified cutoff scores that resulted in similar prevalence rates of severe depression and increased the level of agreement between the scales.

entitled the article "How Can We Use Depression Severity to Guide Treatment Selection When Measures of Depression Categorize Patients Differently?"[14]

It is important to note, however, that the disparities between the scales in the classification of patients into severity groups occurred despite the fact that each of the 3 self-report measures was equally highly correlated with the clinician-rated HDRS. This observation raised a second question: How can 3 scales that each assess the same symptoms of depression be equally valid measures of depression severity yet classify patients differently? We speculated that the approaches used by the scales' developers to identify cutoffs for the severity ranges differentially impacted how broadly the severity groupings were defined. For the QIDS, we could not find a definitive study establishing the severity cutoffs. Several authors refer to the article by Rush et al[16]; however, this study derived QIDS cutoffs corresponding to the definition of remission on the HDRS and did not derive cutoffs corresponding to severity ranges. In a subsequent article, Rush et al[18] identified QIDS scores corresponding to severity ranges and noted the correspondence between QIDS scores and 17-item HDRS scores based on data from their earlier article.[16] Of importance to the issue of severity classification, the 17-item HDRS score used by Rush et al[18] to delineate the lower bound of the severe range was 18, a score that is lower than the usual 17-item HDRS score indicating severe depression.[9,19,20] The cutoff scores for severity ranges on the PHQ-9 were chosen to make them easier for clinicians to recall.[17] The authors also noted that alternative cutoffs did not increase the association between the PHQ-9 severity groupings and indices of construct validity. The severity ranges on the CUDOS were derived empirically.[15] A large sample of psychiatric outpatients completed the scale and were rated on the Clinical Global Impressions-Severity of illness scale (CGI-S).[21] The mean and standard deviation of CUDOS scores were computed for each CGI-S rating, and these values, along with "clinical experience," were used to establish the range of scores for the severity descriptors.

Thus, the developers of these 3 measures used different approaches toward establishing cutoff scores for severity groupings, and, not surprisingly, this resulted in marked differences in the broadness by which severe depression was defined.

There is no consensus in the field as to a preferred self-report measure of depression severity; therefore, it is likely that different scales will continue to be used by both researchers and clinicians. It would be desirable to calibrate the measures against the same criterion standard and establish a cutoff point on each of the scales for severe depression that identifies a similar prevalence of severe depression in depressed outpatients. The goal of the present report from the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project is to derive cutoff scores on each of these 3 measures that will allow for the identification of more comparable groups of patients.

METHOD

As part of an ongoing study of the validity of a new measure to assess remission from depression, from June 2010 to January 2013, 353 outpatients with a principal diagnosis of *DSM-IV* MDD who presented for treatment or who were in ongoing treatment and had their medication changed due to lack of efficacy completed the CUDOS, PHQ-9, and QIDS at the initiation of treatment and were evaluated with the 17-item HDRS by raters blind to the completion of the self-report scales. The sample included 114 men (32.3%) and 239 women (67.7%) who ranged in age from 18 to 84 years (mean = 42.4, SD = 14.3). Approximately two-fifths of the subjects were married (40.5%, n = 143); the remainder were single (27.5%, n = 97), divorced (12.5%, n = 44), separated (6.5%, n = 23), widowed (2.3%, n = 8), or living with someone as if in a marital relationship (10.8%, n = 38). More than half of the patients attended school beyond high school (55.5%, n = 196), though only one-third graduated from a 4-year college (37.1%, n = 131). The racial composition of the sample was 80.5% (n = 284) white, 7.6% (n = 27) black, 7.6% (n = 27) Hispanic, 1.7% (n = 6) Asian, and 2.5% (n = 9) other. The Rhode Island Hospital institutional review committee approved the research protocol, and all patients provided informed, written consent.

The CUDOS contains 16 items assessing all of the *DSM-IV* inclusion criteria for MDD. The respondent is instructed to rate the symptom items on a 5-point ordinal scale indicating "how well the item describes you during the past week, including today" (0 = not at all true/0 days; 1 = rarely true/1–2 days; 2 = sometimes true/3–4 days; 3 = usually true/5–6 days; 4 = almost always true/every day). Compound *DSM-IV* symptom criteria referring to more than 1 construct (eg, problems concentrating or making decisions; insomnia or hypersomnia) were subdivided into their respective components, and a CUDOS item was written for each component. Total scores range from 0 to 64. In the original study[15] of the scale's validity, score ranges were empirically derived corresponding to depression severity categories: no depression, 0–10; minimal depression, 11–20; mild depression, 21–30; moderate depression, 31–45; and severe depression, 46 and above.

Similar to the CUDOS, the QIDS uses 16 items to assess the *DSM-IV* symptom criteria. However, the format of the 2 questionnaires differs. On the QIDS, each symptom is assessed by a group of 4 statements, and the respondent selects the item that best describes how he or she has been feeling. Not every item contributes to the total score. In scoring the QIDS, the highest score is used of the 4 items assessing sleep disturbance (initial, middle, or terminal insomnia or hypersomnia), the 2 items assessing psychomotor disturbance (agitation, retardation), and the 4 items assessing appetite and weight disturbance. Total scores on the scale range from 0 to 27, and the recommended severity score ranges are no depression, 0–5; mild depression, 6–10; moderate depression, 11–15; severe depression, 16–20; and very severe depression, 21–27.[18]

The PHQ-9 contains 9 items corresponding to the *DSM-IV* MDD criteria. Unlike the CUDOS and QIDS, the PHQ-9 assesses compound symptom criteria with a single item. For example, the PHQ-9 assesses insomnia and hypersomnia, and reduced or increased appetite, with a single item. The respondent is instructed to rate the symptom items on a 4-point ordinal scale indicating how often they have been bothered by the symptom over the past 2 weeks (0 = not at all, 1 = several days, 2 = more than half the days, 3 = nearly every day). Total scores on the scale range from 0 to 27, and recommended severity score ranges are no depression, 0–4; mild depression, 5–9; moderate depression, 10–14; moderately severe depression, 15–19; and severe depression, 20–27.[17]

The HDRS is the most commonly used clinician rating scale for depression.[12] We previously reported high reliability in rating the HDRS (intraclass correlation coefficient = 0.97).[22] The cutoff scores to identify severity groups on the 17-item HDRS have varied. Experts on the treatment of severe depression have generally been consistent in recommending a cutoff of 25.[9,19,20] Studies comparing HDRS scores to clinical global severity ratings have validated this cutoff for severe depression.[23,24]

The order of administration of the scales varied in a nonsystematic manner. Some patients completed the self-report scales before the HDRS interview, and some completed them afterward. We did not record this information.

Statistical Analysis

Each of the 3 scales subdivides patients into 5 severity categories, though they do so in different ways. The CUDOS has an extra category at the lower end of severity in which it distinguishes between the absence of clinically significant depression and minimal depression. In contrast, the QIDS and PHQ-9 have an extra category at the severe end of the severity continuum. The PHQ-9 distinguishes between moderately severe and severe depression, whereas the QIDS distinguishes between severe depression and very severe depression. Consistent with our prior report,[14] for the QIDS we combined the 2 highest groups (severe and very severe) into the severe group. Similarly, for the PHQ-9 we combined the 2 highest groups (moderately severe and severe) into the severe group.

Table 1. Concordance Among Depression Measures in Identifying Severe Depression Based on the Scale Developers' Recommended Cutoffs for Severe Depression

	HDRS		CUDOS		PHQ-9	
Scale	κ	% Agreement	κ	% Agreement	κ	% Agreement
CUDOS	0.31	78.4				
PHQ-9	0.21	53.3	0.14	46.7		
QIDS	0.32	67.4	0.31	67.1	0.48	72.2

Abbreviations: CUDOS = Clinically Useful Depression Outcome Scale, HDRS = Hamilton Depression Rating Scale, PHQ-9 = Patient Health Questionnaire, QIDS = Quick Inventory of Depressive Symptomatology.

The goal of the analyses was to identify the cutoff point on each of the self-report scales that would identify a prevalence of severe depression that was similar to that identified by the HDRS (hereafter referred to as the calibration cutoff). Thus, our focus was on maximizing similarity in the prevalence rates, not on maximizing the level of agreement in classification between each measure and the HDRS. If the goal was to maximize agreement in classification, then base rates would be considered as well as a relative balance of sensitivity and specificity. Nonetheless, we also report the cutoff points that maximized the chance-corrected level of agreement between the HDRS and each of the self-report scales.

We computed overall level of agreement and chance-corrected level of agreement (κ) between the measures based on the scale developers' recommended cutoffs and the calibration cutoff. According to the guidelines by Landis and Koch,[25] κ coefficients of 0.20 to 0.39 reflect fair agreement; 0.40 to 0.59, moderate agreement; 0.60 to 0.79, substantial agreement; and 0.80 and above, almost perfect agreement.

RESULTS

Six patients failed to complete at least 1 of the 5 depression scales. A total of 347 patients completed all depression scales and is the basis of all subsequent analyses.

For the HDRS and CUDOS, the mean (SD) score fell in the moderate range (HDRS: 20.1 [6.0]; CUDOS: 34.4 [10.9]). In contrast, the mean (SD) scores on the PHQ-9 (16.8 [5.5]) and QIDS (15.5 [4.5]) fell into the moderately severe and severe ranges, respectively. A minority of patients had severe depression according to the HDRS (23.1%, n = 80) and CUDOS (15.3%, n = 53). In contrast, the majority of patients were severely depressed according to the PHQ-9 (67.4%, n = 234) and QIDS (46.4%, n = 161). The level of agreement between the 3 self-report scales and the HDRS was fair (overall level of agreement: mean = 64.4%, κ = 0.28), as was the level of agreement between the pairs of self-report scales (overall level of agreement: mean = 62.0%, κ = 0.31) (Table 1).

Table 2 shows the percentage of patients scoring above each cutoff point on each scale. The cutoffs that identified a prevalence of severe depression most similar to the prevalence based on the HDRS were 43 for the CUDOS (prevalence = 22.2%), 22 for the PHQ-9 (prevalence = 22.8%), and 19 for the QIDS (prevalence = 26.5%). Based on these cutoffs, the level of agreement between the self-report scales

Table 2. Percentage of Depressed Outpatients Scoring at the Same Level on 3 Self-Report Depression Scales

Cutoff Score[a]	Clinically Useful Depression Outcome Scale (CUDOS)	Patient Health Questionnaire (PHQ-9)	Quick Inventory of Depressive Symptomatology (QIDS)
≥ 10	98.8	89.6	90.5
≥ 11	98.6	83.3	86.5
≥ 12	98.0	80.1	82.4
≥ 13	97.4	76.7	74.9
≥ 14	96.8	71.2	67.7
≥ 15	96.3	67.4	58.2
≥ 16	94.8	62.5	46.4
≥ 17	93.9	57.1	38.3
≥ 18	93.4	49.0	32.9
≥ 19	91.9	41.2	**26.5**
≥ 20	91.4	33.4	19.6
≥ 21	89.9	27.7	14.4
≥ 22	87.9	**22.8**	9.8
≥ 23	85.3	15.9	7.2
≥ 24	83.0	10.1	4.0
≥ 25	81.0	6.1	2.0
≥ 26	78.4	4.3	0.6
≥ 27	76.9	2.6	0.3
≥ 28	74.6	0.0	0.0
≥ 29	71.2		
≥ 30	68.3		
≥ 31	65.1		
≥ 32	61.4		
≥ 33	58.2		
≥ 34	55.6		
≥ 35	51.0		
≥ 36	48.1		
≥ 37	43.8		
≥ 38	39.8		
≥ 39	35.7		
≥ 40	32.9		
≥ 41	30.0		
≥ 42	25.6		
≥ 43	**22.2**		
≥ 44	20.7		
≥ 45	17.6		
≥ 46	15.3		
≥ 47	13.8		
≥ 48	11.2		
≥ 49	9.2		
≥ 50	8.6		

[a]The table should be read as follows: 98.8% of the patients scored 10 or higher on the CUDOS, 89.6% on the PHQ-9, and 90.5% on the QIDS. The value closest to the 23.1% prevalence rate of severe depression according to the Hamilton Depression Rating Scale is shown in boldface.

Table 3. Concordance Among Depression Measures in Identifying Severe Depression Based on Cutoffs Calibrated to the HDRS Severity Prevalence Rate

	HDRS		CUDOS		PHQ-9	
Scale	κ	% Agreement	κ	% Agreement	κ	% Agreement
CUDOS	0.42	79.7				
PHQ-9	0.37	77.8	0.53	83.4		
QIDS	0.41	78.1	0.57	84.2	0.55	83.1

Abbreviations: CUDOS = Clinically Useful Depression Outcome Scale, HDRS = Hamilton Depression Rating Scale, PHQ-9 = Patient Health Questionnaire, QIDS = Quick Inventory of Depressive Symptomatology.

Table 4. Proposed Cutoff Scores Among Depression Measures in Identifying Severe Depression Based on the Hamilton Depression Rating Scale Severity Prevalence Rates

Scale	Cutoff	% Severe
Hamilton Depression Rating Scale	25	22.9
Clinically Useful Depression Outcome Scale	43	22.2
Patient Health Questionnaire	22	22.8
Quick Inventory of Depressive Symptoms	19	26.5

and HDRS in identifying severe depression was higher than it was based on the scale developers' cutoffs (overall level of agreement: mean = 78.5%, κ = 0.40) (Table 3). Likewise, the level of agreement between the self-report scales increased (overall level of agreement: mean = 83.5%, κ = 0.55). The cutoff score maximizing the chance-corrected level of agreement between the self-report scales and the HDRS was similar or identical to the cutoff that maximized the concordance of prevalence rates (CUDOS, cutoff of 42, κ = 0.45; PHQ-9, cutoff of 21, κ = 0.41; QIDS, cutoff of 19, κ = 0.41).

DISCUSSION

Official treatment guidelines for depression suggest treatment options on the basis of severity distinctions.[6,11] All treatment guidelines recommend pharmacotherapy as the treatment of choice for severe depression. However, it is difficult to implement and evaluate these recommendations in clinical practice if clinicians use different scales that vary in how broadly the severe depression category is defined.

To the best of our knowledge, this is the first study to have calibrated 3 self-report scales that were each designed to measure the severity of depression based on the *DSM-IV* symptom criteria. The developers of the QIDS derived the cutoff for severe depression based on the association between the scale and the 17-item HDRS. However, they used a cutoff of 18 on the 17-item HDRS, which is a much lower score than what is typically used (and empirically validated) to define severe depression. It is therefore not surprising that the QIDS cutoff of 16 for severe depression is very broad. In a sample of primary care patients with depression, Cameron et al[26] similarly found that the QIDS was overinclusive in identifying severe depression. The results of the present study suggest that the cutoff on the QIDS for severe depression should be raised from its currently recommended value of 16 (Table 4).

The developers of the PHQ-9 did not base the cutoff for severe depression on a statistical analysis of the type that is typically conducted to derive cutoff scores. Similar to the results of our study, other studies have found that the PHQ-9 was overinclusive in identifying severe depression.[27–29] Our findings suggest that the recommended cutoff for severe depression should be raised (Table 4).

In developing the CUDOS, our group derived a cutoff for severe depression by comparing CUDOS scores to the CGI.[15] The findings of the present study suggest that the previously recommended cutoff of 46 for severe depression should be lowered (Table 4).

The calibration cutoffs derived in the present study produced a similar prevalence of severe depression on the 3 self-report scales, and the level of agreement between the scales in the classification of severe depression increased. Our analyses focused on identifying cutoffs that resulted in similar prevalence rates of severe depression. When we derived cutoff scores based on maximizing the agreement with the HDRS in classifying severe depression, the cutoffs were identical for 1 scale and differed by 1 point for the other 2 scales. One could reasonably argue that it would be more appropriate to select a cutoff that maximized agreement between the self-report scales and the HDRS because the purpose of identifying such cutoffs is to most validly determine which patients seen in clinical practice require pharmacologic treatment in preference to psychotherapy. However, a limitation with the recommendations in official treatment guidelines is that they are based on clinician-rated instruments, such as the HDRS (which does not directly correspond to the *DSM-IV* symptom criteria), that have been applied to patients in efficacy studies (who are not representative of patients in routine clinical practice). *DSM-IV*–based self-report measures do not perfectly agree with clinician measures, such as the HDRS, in identifying severe depression (the postcalibration agreement rates between the 3 self-report scales and HDRS were slightly lower than 80%, and the κ coefficients were all below 0.50). Whether self-report scales would also be valid in predicting differential treatment response has not been demonstrated. Thus, while the results of this calibration study increase the concordance between the measures in identifying severe depression and should make it easier to compare results across studies that use different measures, it nonetheless will be important for future research to determine whether these self-report scales predict differential treatment response. The need for such research on the predictive power of self-report scales of the type included in the present study is particularly salient for clinical purposes because such scales are more likely than clinician-rated instruments to be used to measure the severity of depressive symptoms in routine clinical practice.

As we have noted elsewhere,[30] it seems unreasonable to us that, in a series of studies that used the PHQ-9 to assess depression severity, more than half of depressed outpatients had severe depression. We therefore expressed concern that reliance on a self-report scale that tends to be overinclusive in classifying depressed patients as severe could result in the overprescription of medication and underutilization of psychotherapy. Consequently, in the absence of research demonstrating that, in routine clinical practice, self-report scales validly identify a subset of patients with MDD with severe depression who should preferentially be treated with medication, the emphasis of the current report has been on identifying cutoff points that maximize similarity in prevalence rates of severe depression rather than identifying cutoff points that maximize agreement with the HDRS.

The criterion standard for identifying severe depression in the current study was a score of 25 and higher on the 17-item HDRS. To be sure, there is not a consensus in the field to use this cutoff to define severe depression. In the APA's *Handbook of Psychiatric Measures*,[31] a cutoff of 19 was recommended to define severe depression. Two studies were cited in support of this cutoff. One was a study[32] examining the validity of deriving an HDRS equivalent score on the Schedule for Affective Disorders and Schizophrenia. In fact, this study did not attempt to determine the cutoff scores on the HDRS indicating grades of severity. Rather, when examining the agreement between the extracted and original HDRS in classifying patients into severity categories, the authors used a cutoff of 25 to indicate severe depression (and a cutoff of 18 to distinguish mild and moderate depression). The second study,[33] cited as evidence for using a cutoff of 19 to indicate severe depression, examined the association between HDRS scores and global ratings of severity in 59 depressed inpatients. The authors did not derive (or recommend) cutoff scores corresponding to severity levels. In Figure 2 of their article, the authors graphed the mean HDRS scores for patients rated at different levels of severity. Visual inspection of this figure suggests that very severe depression corresponded to a mean HDRS score of approximately 29, and severe depression corresponded to a mean HDRS score of 21. If these groups were combined, the mean HDRS score for the severe category would be approximately 25. Thus, it is unclear why a cutoff of 19 was recommended in the APA *Handbook* to identify severe depression. We are aware of only 2 other small studies comparing HDRS scores to clinical global severity ratings. Knesevich et al[23] evaluated a sample of 26 outpatients, 9 of whom were rated in the severe range. Visual inspection of the figure plotting the distribution of scores suggests that the median score for these patients was 24. Muller et al[24] evaluated 85 depressed inpatients, 26 of whom were rated severe. They conducted a receiver operating characteristic analysis to determine the optimal cutoff score on the HDRS to indicate severe depression and found that a cutoff of 25 provided the best balance of sensitivity and specificity. These findings, together with the recommendations of experts in the treatment of severe depression,[9,19,20] support our choice of a cutoff of 25 on the 17-item HDRS to delineate severe depression.

As noted above, the recommendations from treatment guidelines regarding treatment selection are based on controlled efficacy treatment trials, and the generalizability of these efficacy studies to patients seen in routine clinical practice has been challenged.[34] The patients in the present study did not pass through the inclusion and exclusion criteria filters that are typical of these studies. Thus, a question can be raised regarding merits of the goal of calibrating the self-report scales to an HDRS cutoff that is based on efficacy studies of uncertain generalizability. Replication of the present findings in a sample participating in an efficacy trial is warranted.

A limitation of this study is that it was conducted in a single clinical practice in which the majority of the patients were white and female and had health insurance. Replication in samples with different demographic characteristics is warranted. However, the generalizability of the findings is enhanced by the lack of inclusion and exclusion criteria to select patients. Severity was defined according to scores on symptom severity measures. Other methods to determine severity, such as hospitalization, the presence of melancholic features, and the level of functional impairment, have also been used as indicators of severity.[9] There are advantages and disadvantages to each of these approaches to determine the severity of depression, though scores on standardized rating scales have been the most commonly used index of severity.

The present study focused on cutoffs to identify severe depression and did not attempt to derive score ranges for mild and moderate depression. We focused on severe depression because the APA treatment guidelines identify severe depression as requiring a particular treatment modality (ie, pharmacologic), whereas mild and moderate depression can be addressed pharmacologically or psychotherapeutically. Also, the patients in the study were in a depressive episode at the time of the assessment; therefore, we would be unable to identify the lower bound of the mild depression group.

The goal of the present study was to calibrate the 3 self-report scales against a criterion standard, clinician-rated instrument. The content of the HDRS does not perfectly match the content of the *DSM-IV*–based self-report scales. Thus, the modest levels of agreement between the self-report measures and the HDRS are not surprising. Higher levels of agreement would be expected between measures of identical content, and this might account for the postcalibration higher agreement levels among the self-report scales than between the self-report scales and the HDRS.

Author affiliations: Department of Psychiatry and Human Behavior, Brown Medical School, and Department of Psychiatry, Rhode Island Hospital, Providence.

Potential conflicts of interest: None reported.

Funding/support: The research was supported, in part, by Eli Lilly USA, LLC.
Role of the sponsor: The sponsor had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

REFERENCES

1. Khan A, Leventhal RM, Khan SR, et al. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Psychopharmacol*. 2002;22(1):40–45.
2. Elkin I, Gibbons RD, Shea MT, et al. Initial severity and differential treatment outcome in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *J Consult Clin Psychol*. 1995;63(5):841–847.
3. Fournier JC, DeRubeis RJ, Hollon SD, et al. Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA*. 2010;303(1):47–53.
4. Driessen E, Cuijpers P, Hollon SD, et al. Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? a meta-analysis. *J Consult Clin Psychol*. 2010;78(5):668–680.
5. Wiles NJ, Mulligan J, Peters TJ, et al. Severity of depression and response to antidepressants: GENPOD randomised controlled trial. *Br J Psychiatry*. 2012; 200(2):130–136.
6. American Psychiatric Association. *Practice Guideline for the Treatment of Patients With Major Depressive Disorder*. 3rd ed. Washington, DC: American Psychiatric Association; 2010.
7. Kilts CD, Wade AG, Andersen HF, et al. Baseline severity of depression predicts antidepressant drug response relative to escitalopram. *Expert Opin Pharmacother*. 2009;10(6):927–936.
8. Schmitt AB, Bauer M, Volz HP, et al. Differential effects of venlafaxine in the treatment of major depressive disorder according to baseline severity. *Eur Arch Psychiatry Clin Neurosci*. 2009;259(6):329–339.
9. Schatzberg AF. Antidepressant effectiveness in severe depression and melancholia. *J Clin Psychiatry*. 1999;60(suppl 4):14–21, discussion 22.
10. van der Lem R, van der Wee NJ, van Veen T, et al. The generalizability of antidepressant efficacy trials to routine psychiatric out-patient practice. *Psychol Med*. 2011;41(7):1353–1363.
11. National Collaborating Center for Mental Health. Depression: the treatment and management of depression in adults. London, England: National Institute for Health and Clinical Excellence; 2009: 64. http://publications.nice.org.uk/depression-in-adults-cg90. Updated October 2009. Accessed July 8, 2013.
12. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23(1):56–62.
13. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry*. 1979;134(4):382–389.
14. Zimmerman M, Martinez JH, Friedman M, et al. How can we use depression severity to guide treatment selection when measures of depression categorize patients differently? *J Clin Psychiatry*. 2012;73(10):1287–1291.
15. Zimmerman M, Chelminski I, McGlinchey JB, et al. A clinically useful depression outcome scale. *Compr Psychiatry*. 2008;49(2):131–140.
16. Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry*. 2003;54(5):573–583.
17. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16(9):606–613.
18. Rush AJ, Bernstein IH, Trivedi MH, et al. An evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: a sequenced treatment alternatives to relieve depression trial report. *Biol Psychiatry*. 2006;59(6):493–501.
19. Hirschfeld RM. Efficacy of SSRIs and newer antidepressants in severe depression: comparison with TCAs. *J Clin Psychiatry*. 1999;60(5):326–335.
20. Montgomery SA, Lecrubier Y. Is severe depression a separate indication? ECNP Consensus Meeting September 20, 1996, Amsterdam. European College of Neuropsychopharmacology. *Eur Neuropsychopharmacol*. 1999; 9(3):259–264.
21. Guy W. *ECDEU Assessment Manual for Psychopharmacology*. US Department of Health, Education, and Welfare publication (ADM) 76-338. Rockville, MD: National Institute of Mental Health; 1976:218–222.
22. Zimmerman M, Posternak MA, Chelminski I. Derivation of a definition of remission on the Montgomery-Asberg depression rating scale corresponding to the definition of remission on the Hamilton rating scale for depression. *J Psychiatr Res*. 2004;38(6):577–582.
23. Knesevich JW, Biggs JT, Clayton PJ, et al. Validity of the Hamilton Rating Scale for depression. *Br J Psychiatry*. 1977;131(1):49–52.
24. Müller MJ, Himmerich H, Kienzle B, et al. Differentiating moderate and severe depression using the Montgomery-Asberg depression rating scale (MADRS). *J Affect Disord*. 2003;77(3):255–260.
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
26. Cameron IM, Crawford JR, Cardy AH, et al. Psychometric properties of the Quick Inventory of Depressive Symptomatology (QIDS-SR) in UK primary care. *J Psychiatr Res*. 2013;47(5):592–598.
27. Cameron IM, Cardy A, Crawford JR, et al. Measuring depression severity in general practice: discriminatory performance of the PHQ-9, HADS-D, and BDI-II. *Br J Gen Pract*. 2011;61(588):e419–e426.
28. Cameron IM, Crawford JR, Lawton K, et al. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *Br J Gen Pract*. 2008;58(546):32–36.
29. Hansson M, Chotai J, Nordstöm A, et al. Comparison of two self-rating scales to detect depression: HADS and PHQ-9. *Br J Gen Pract*. 2009; 59(566):e283–e288.
30. Zimmerman M. Symptom severity and guideline-based treatment recommendations for depressed patients: implications of *DSM-5*'s potential recommendation of the PHQ-9 as the measure of choice for depression severity. *Psychother Psychosom*. 2012;81(6):329–332.
31. Rush AJ, First MB, Blacker D. *Handbook of Psychiatric Measures*. 2nd ed. Washington, DC: American Psychiatric Publishing, Inc; 2008.
32. Endicott J, Cohen J, Nee J, et al. Hamilton Depression Rating Scale: extracted from Regular and Change Versions of the Schedule for Affective Disorders and Schizophrenia. *Arch Gen Psychiatry*. 1981;38(1):98–103.
33. Kearns NP, Cruickshank CA, McGuigan KJ, et al. A comparison of depression rating scales. *Br J Psychiatry*. 1982;141(1):45–49.
34. Zimmerman M, Mattia JI, Posternak MA. Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *Am J Psychiatry*. 2002;159(3):469–473.