Available online at www.sciencedirect.com
ScienceDirect
Comprehensive Psychiatry 49 (2008) 131 – 140
Comprehensive PSYCHIATRY
www.elsevier.com/locate/comppsych

A clinically useful depression outcome scale

Mark Zimmerman*, Iwona Chelminski, Joseph B. McGlinchey, Michael A. Posternak
Department of Psychiatry and Human Behavior, Brown University School of Medicine, Providence, RI 02905, USA

Abstract

If the optimal delivery of mental health treatment ultimately depends on examining outcome, then precise, reliable, valid, informative, and user-friendly measurement is the key to evaluating the quality and efficiency of care in clinical practice. Self-report questionnaires are a cost-effective option because they are inexpensive in terms of professional time needed for administration, and they correlate highly with clinician ratings. In the present report from the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project, we describe the reliability and validity of the Clinically Useful Depression Outcome Scale (CUDOS). The CUDOS was designed to be brief (completed in less than 3 minutes), quickly scored (in less than 15 seconds), clinically useful (fully covering the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* symptoms of major depressive disorder and dysthymic disorder), reliable, valid, and sensitive to change. We studied the CUDOS in more than 1400 psychiatric outpatients and found that the scale had high internal consistency and test-retest reliability. The CUDOS was more highly correlated with another self-report measure of depression than with measures of anxiety, substance use problems, eating disorders, and somatization, thereby supporting the convergent and discriminant validity of the scale. The CUDOS was also highly correlated with interviewer ratings of the severity of depression, and CUDOS scores were significantly different in depressed patients with mild, moderate, and severe levels of depression. The CUDOS was a valid measure of symptom change. Finally, the CUDOS was significantly associated with a diagnosis of major depressive disorder. Thus, the results of this large validation study of the CUDOS shows that it is a reliable and valid measure of depression that is feasible to incorporate into routine clinical practice.

1. Introduction

To determine the impact of treatment it is necessary to evaluate outcome. In mental health clinical practice, this typically is based on unstructured interactions that yield unquantified judgments of progress. This is at variance with other areas of medical care in which outcome is determined, in part, on the change of a numerical value. Body temperature, blood pressure, cholesterol values, blood sugar levels, cardiac ejection fraction, and white blood cell counts are examples of quantifiable variables that are used to evaluate treatment progress. In the mental health field, standardized, quantifiable outcome measures exist for most major psychiatric disorders, yet they are rarely used in routine clinical practice [1].

In the current health care environment, which is concerned with cost containment, treatment efficiency is assuming increasing importance because there is an interest in the relationship between the cost and results of providing services. Cost and outcome are relevant issues to practicing clinicians, clinic administrators, and third party payers, and addressing fiscal concerns can impact on the delivery of appropriate clinical services in routine clinical settings. To determine treatment efficiency in clinical practice, it would require treatment providers to adopt some of the techniques that are standard in research treatment outcome investigations.

If the optimal delivery of mental health treatment ultimately depends on examining outcome, then precise, reliable, valid, informative, and user-friendly measurement is critical to evaluating the quality and efficiency of care in clinical practice. Clinicians are already overburdened with paperwork, and adding to this load by requiring repeated detailed evaluations with such instruments as the Hamilton Rating Scale for Depression (HAMD) [2] is unlikely to meet

* Corresponding author.
E-mail address: mzimmerman@lifespan.org (M. Zimmerman).

0010-440X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.comppsych.2007.10.006

with success. Self-report questionnaires are a cost-effective option because they are inexpensive in terms of professional time needed for administration and they correlate highly with clinician ratings. To be sure, there are also limitations with self-report questionnaires, such as response-set biases, and their use may be limited by the readability of the scale and literacy of the respondent. However, self-report scales are free of clinician bias and are therefore free from clinician overestimation of patient improvement (which might occur when there is incentive to document treatment success).

Three consumers should be considered in the construction of a self-administered outcome questionnaire to be used in routine clinical practice: the patient, the clinician, and the administrator. Patients should find the measure user-friendly and the directions easy to follow. The questions should be understandable and relevant to the patient's problem. The scale should be brief, taking no more than 2 to 3 minutes to complete, so that upon routine administration at follow-up visits, patients are not inconvenienced by the need to come for their appointment 10 to 15 minutes early in order to complete the measure. This would make it feasible to have the scale completed at each follow-up visit in the same way that blood pressure, cholesterol levels, and weight are routinely assessed in primary care settings for patients being treated for hypertension, hypercholesterolemia, and obesity.

The instrument should provide clinicians with clinically useful information and improve the efficiency of conducting their clinical evaluation; thus, the measure should have practical value to the practicing clinician. Of course, clinicians need to be able to trust the information provided by any instrument they use. Consequently, outcome measures of any kind should have a sound basis in psychometrics, demonstrating good reliability, validity, and sensitivity to change. Clinicians and clinics should also find the instrument user-friendly; it should be easy to administer and score with minimal training.

Clinic administrators likewise want measures to be both reliable and valid. To successfully implement an outcomes assessment program, administrators want measures to have high patient and clinician acceptance. Administrators are also concerned about the cost of an instrument, from the perspective of both the purchase price and the cost of labor to score the scale. Thus, an outcome measure, or outcome assessment program, should be inexpensive to purchase and implement.

Finally, we believe that any instrument constructed for use in clinical settings should meet scientific standards for publication in peer-reviewed journals. It is important that a new measure stand up to critical scientific review and be published in the scientific arena so that other investigators may further examine its properties.

During the past decade, we have established and have been conducting the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project. One of the goals of the MIDAS project is to develop instruments for use in routine clinical practice. Previously, we have described the reliability and validity of a broad-based self-report scale for psychiatric screening [3-5]; a clinician-rated outcome measure for depression [6]; and single-item self-report indices of depression symptom severity, psychosocial impairment, and quality of life [7]. In the present report from the MIDAS project, we describe the reliability and validity of the Clinically Useful Depression Outcome Scale (CUDOS). The CUDOS was designed to be brief (completed in less than 3 minutes), quickly scored (in less than 15 seconds), clinically useful (fully covering the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* [*DSM-IV*] symptoms of major depressive disorder [MDD] and dysthymic disorder), reliable, valid, and sensitive to change. Certainly, there is no shortage of self-report depression scales [8]. However, other questionnaires are either too long [9-11], lack adequate coverage of the *DSM-IV* diagnostic criteria [12,13], are expensive to purchase [9], or are somewhat complicated to score [12]. These factors reduce their appeal as outcome tools for use in routine clinical practice. Table 1 lists 15 desirable properties of a depression outcome scale. The CUDOS was designed to possess all 15 characteristics of a depression outcome scale that could be routinely incorporated into clinical practice.

2. Methods

Individuals presenting for an intake evaluation at the Rhode Island Hospital Department of Psychiatry outpatient practice were asked to complete the CUDOS as part of their initial paperwork. Because we were planning to test the CUDOS' validity by examining its relationship with psychiatric diagnoses, the diagnosticians were kept blind to the subjects' responses on the measure. The Rhode Island Hospital institutional review committee approved the research protocol, and all patients provided informed, written consent.

One thousand four hundred seventy-five psychiatric outpatients completed the CUDOS at the time of their intake

Table 1
Desirable features of a self-report depression outcome scale

1. Brief
2. Acceptable to patients
3. Covers all *DSM-IV* diagnostic criteria for MDD
4. Reliable (internal consistency and test-retest reliability)
5. Convergent validity
6. Discriminant validity
7. Indicator of symptom severity
8. Indicator of remission status
9. Case-finding capability as a screening instrument
10. Assesses psychosocial function
11. Assesses quality of life
12. Assesses suicidal thoughts
13. Sensitive to change
14. Easy to score
15. Inexpensive

appointment. The group included 589 men (39.9%) and 886 women (60.1%) who ranged in age from 18 to 85 years (mean = 38.4, SD = 13.0). About two fifths of the subjects were married (41.4%, n = 611); the remainder were single (31.0%, n = 457), divorced (15.3%, n = 226), separated (4.8%, n = 71), widowed (1.9%, n = 28), or living with someone as if in a marital relationship (5.6%, n = 82). The educational level achieved by the subjects was as follows: 8.7% (n = 129) did not graduate high school, 63.7% (n = 940) graduated high school or achieved equivalency, and 27.5% (n = 406) graduated college. The racial composition of the sample was 87.1% (n = 1285) white, 4.9% (n = 73) black, 2.8% (n = 41) Hispanic, 0.9% (n = 13) Asian, and 4.3% (n = 63) from another or a combination of the above racial backgrounds.

The CUDOS contains 18 items assessing all of the *DSM-IV* inclusion criteria for MDD and dysthymic disorder as well as psychosocial impairment and quality of life. The 16 symptom items were derived from a larger pool of 27 items. Alternative wordings of items were written, and the psychometric performance of the alternative forms was compared to select the best performing versions of the items. For items that were similar in wording (e.g., My energy level was low. I felt tired and worn out.), we selected for inclusion on the final version the item which was more highly correlated with the total score. Compound *DSM-IV* symptom criteria referring to more than one 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. The individual symptoms assessed by the CUDOS are depressed mood, loss of interest in usual activities, low energy, psychomotor agitation, psychomotor retardation, guilt, worthlessness, thoughts of death, suicidal ideation, impaired concentration, indecisiveness, decreased appetite, increased appetite, insomnia, hypersomnia, and hopelessness. Because of an oversight, the initial version of the CUDOS did not contain the item for psychomotor agitation. However, the correlation between the total scale score with and without the agitation item was very high (r = .99); thus, the data were combined for all patients. The CUDOS also includes items assessing global perception of psychosocial impairment due to depression and overall quality of life. The reliability and validity of these items has been described elsewhere [7].

The respondent is instructed to rate the symptom items on a 5-point Likert 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). A 1-week time frame was preferred to allow for weekly assessments. Although this might compromise the diagnostic validity of the scale compared to the *DSM-IV* definition of major depression, which is based on a 2-week time frame, we thought it was more important to have a measure that could be used for weekly assessments than one that more closely followed the *DSM-IV* diagnostic algorithm. Moreover, the potential increase in false positives by using a 1-week time frame should be counterbalanced by a decrease in false negatives (ie, sensitivity would be increased, and specificity, decreased). A Likert rating of the symptom statements was preferred in order to keep the scale brief. Scales such as the Beck Depression Inventory (BDI) [9], Diagnostic Inventory for Depression [14], and Inventory of Depressive Symptoms [11] assess symptoms with groups of 4 or 5 statements and are thus composed of 80 or more statements. These scales take respondents 5 to 15 minutes to complete, and this was considered too long for regular use in clinical practice in which the scale would be routinely administered at follow-up appointments. Consistent with this, a comparison of the feasibility and acceptability of the CUDOS and the BDI to assess outcome in clinical practice found that significantly more patients indicated that the CUDOS took less time to complete and was less of a burden to complete [15].

All patients were interviewed by a trained diagnostic rater who administered the Structured Clinical Interview for *DSM-IV* (SCID) [16] supplemented with questions from the Schedule for Affective Disorders and Schizophrenia (SADS) [17] assessing the severity of symptoms during the week prior to the evaluation. An extracted HAMD score was derived from the SADS ratings following the algorithm developed by Endicott et al [18]. Patients were also rated on the Clinical Global Index (CGI) of depression severity [19] and *DSM-IV*'s Global Assessment of Functioning scale. Details regarding interviewer training and diagnostic reliability are available in other publications from the MIDAS project [20-22].

To examine the convergent and discriminant validity of the CUDOS, 204 subjects completed a booklet of questionnaires at home that included measures of symptoms related to bulimia (Eating Disorder Inventory–anorexia and bulimia subscales [23]), depression (BDI [24]), social phobia (Brief Fear of Negative Evaluation Scale [25], Fear Questionnaire–social phobia subscale [26]), agoraphobic fears and cognitions (Fear Questionnaire–agoraphobia subscale [26], Social Phobia and Anxiety Inventory–agoraphobia subscale [27]), posttraumatic stress disorder (Posttraumatic Stress Disorder Scale [28]), obsessive-compulsive behavior (Maudsley Obsession-Compulsion Questionnaire [29]), cognitions common in generalized anxiety (Penn State Worry Scale [30]), anxiety symptoms and cognitions common in panic disorder (Beck Anxiety Inventory [31]), alcohol use (Michigan Alcohol Screening Test [32]), drug use (Drug Abuse Screening Test [33]), hypochondriasis (Whitely Index [34]), somatization (Somatic Symptom Index [35,36]), mania (Self-Report Mania Inventory [37]), and psychosis (Symptom Rating Test–psychosis and paranoia subscales [38]). These scales have been widely used, and their reliability and validity well established.

The test-retest reliability of the CUDOS was examined in 2 samples—one collected at the time of the initial evaluation and the other collected from depressed patients in ongoing

treatment. Both samples were of consecutively ascertained patients. Test-retest reliability may vary as a function of when during the course of treatment reliability data is collected. At intake, patients are usually distressed and highly symptomatic and symptom change might occur over short periods of time. Score variation between the initial and repeat administration of a symptom scale may validly reflect true change in clinical status but would be interpreted as indicating scale unreliability. To control for this, the test-retest interval should be brief. In contrast, patients in ongoing treatment are more likely to have stable symptom levels, and accurate estimates of scale reliability might be obtained using a longer retest interval. One hundred seventy-six patients who completed the CUDOS at the time of their first appointment were given the scale at the conclusion of the intake evaluation and asked to mail it back in a preaddressed postage paid envelope. They were told that the purpose of the second administration was to test the performance of the scale, not to question the truthfulness or accuracy of their responses. To minimize the influence of changing state, we examined test-retest reliability for the 176 patients who completed the second administration within a week of the first. A second sample of 33 patients in ongoing treatment completed the CUDOS at the time of their appointment and were asked to complete it again a week later and to return it by mail.

Sensitivity to change was examined in 55 depressed patients who completed the CUDOS a second time 6 weeks after treatment had begun. At the follow-up evaluation, the treating clinicians rated the subjects on the Montgomery-Asberg Depression Rating Scale (MADRS) [39] and the CGI improvement scale [19].

3. Data analyses

We examined 2 types of reliability of the CUDOS test-retest reliability and internal consistency. We examined convergent and discriminant validity [40] by comparing the correlation between the CUDOS and the Beck Depression Inventory with the correlation between the CUDOS and measures of anxiety, substance use, eating disorders, and somatization. Because depression is frequently comorbid with other Axis I disorders, we predicted that the CUDOS would be significantly correlated with measures of nondepressive symptom domains, although the correlation with another measure of depression would be significantly higher.

We examined the diagnostic properties of the CUDOS. Sensitivity refers to a test's ability to correctly identify individuals with the disorder, whereas specificity refers to a test's ability to identify non-ill persons. Sensitivity and specificity provide useful psychometric information about a test; however, the clinically more meaningful conditional probabilities are positive and negative predictive values. These values indicate the probability that an individual is ill or non-ill, given that the test identifies them as ill or non-ill. Accordingly, positive predictive value is the percentage of individuals classified ill by the test who truly are ill, whereas negative predictive value is the percentage of individuals classified not ill by the test who truly are not ill. Unlike sensitivity and specificity, positive and negative values are influenced by the prevalence of the disorder.

Depending on the scale's purpose, cutoff scores might be selected to optimize the sensitivity or specificity of the scale [41,42]. We determined diagnostic performance across the range of cutoff scores by conducting receiver operating curve (ROC) analyses [41]. A ROC curve is a plot of a measure's sensitivity versus one minus specificity at each cutoff score. The area under this curve is the evaluative measure, which can range from .5 (random performance) to 1.0 (perfect performance).

4. Results

4.1. Internal consistency and test-retest reliability of the CUDOS

The data in Table 2 shows the diagnostic characteristics of the 1475 patients who completed the CUDOS at their initial appointment. The most frequent *DSM-IV* diagnoses were MDD (42.4%), social phobia (26.0%), generalized anxiety disorder (18.8%), and panic disorder (17.2%).

Internal consistency coefficients were computed separately for the 1475 patients who completed the scale at intake and 100 depressed patients who participated in an acceptance and feasibility study who completed it during a follow-up appointment [15]. The CUDOS demonstrated excellent internal consistency in both samples (Cronbach α at intake = .90; Cronbach α at follow-up = .90). The data in Table 3 shows the correlation between each item and the

Table 2
DSM-IV Axis I diagnoses of 1475 psychiatric outpatients

DSM-IV diagnosis	n	%
MDD	625	42.4
Bipolar disorder	84	5.7
Dysthymic disorder	125	8.5
Generalized anxiety disorder	273	18.5
Panic disorder	254	17.2
Social phobia	384	26.0
Specific phobia	140	9.5
Obsessive-compulsive disorder	98	6.6
Posttraumatic stress disorder	185	12.5
Adjustment disorder	90	6.1
Schizophrenia	6	0.4
Eating disorder	102	6.9
Alcohol abuse/dependence	133	9.0
Drug abuse/dependence	72	4.9
Somatoform disorder	103	7.0
Attention deficit disorder	75	5.1
Impulse control disorder	87	5.9

Individuals could be given more than 1 diagnosis.

Table 3
Item-total correlations and test-retest reliability of individual CUDOS items at baseline and follow-up

CUDOS Items	Item-total correlations		Test-retest reliability	
	Baseline	Follow-up	Baseline	Follow-up
	(n = 1,475)	(n = 100)	(n = 176)	(n = 33)
1. Depressed mood	.75	.75	.79	.92
2. Decreased interest in usual activities	.77	.82	.76	.90
3. Decreased appetite	.50	.37	.77	.87
4. Increased appetite	.34	.14	.53	.75
5. Insomnia	.46	.56	.77	.71
6. Hypersomnia	.44	.27	.67	.46
7. Psychomotor agitation	.66	.34	missing	.76
8. Psychomotor retardation	.74	.56	.66	.84
9. Decreased energy	.72	.73	.76	.83
10. Guilt	.68	.77	.65	.76
11. Worthlessness	.75	.76	.80	.91
12. Decreased concentration	.72	.80	.75	.94
13. Indecisiveness	.76	.72	.75	.72
14. Thoughts about death	.66	.48	.81	.80
15. Suicidal ideation	.53	.37	.77	.99
16. Hopelessness	.73	.66	.78	.87

All correlations are significant at $P < .001$.

total scale score. All item-scale correlations at baseline and follow-up were significant (mean at intake = .64; mean at follow-up = .57).

The test-retest reliability of the CUDOS was examined in 176 subjects at baseline and 33 subjects during follow-up. At both time points, the test-retest reliability of the total scale was high (r = .92 and .95, respectively), and the test-retest reliability of each item was significant (mean at intake = .73; mean at follow-up = .81) (Table 3).

4.2. Discriminant and convergent validity of the CUDOS

Two hundred subjects completed a package of questionnaires at home an average of 1.2 days (SD = 16.9) after the intake evaluation. The data in Table 4 show that the CUDOS was more highly correlated with the BDI than with measures of the other symptom domains. Moreover, the CUDOS was nearly as highly correlated with the extracted HAMD and the CGI, clinician ratings of the severity of depressive symptoms, as with the self-rated BDI.

4.3. The ability of the CUDOS to discriminate between levels of depression severity

The ability of the CUDOS to discriminate between different levels of depression severity was examined with an analysis of variance on the CGI depression severity rating. Because only 5 of the 1475 patients were rated at the highest level of severity (extreme depression), the 2 highest rating levels were combined. The overall analysis of variance was significant (F = 351.0; df = 4,1470; $P < .001$). As Fig. 1 illustrates, the CUDOS scores increased with increasing global severity ratings (nondepressed, 13.0 ± 10.4; minimal, 20.7 ± 10.4; mild, 27.9 ± 9.9; moderate 37.5 ± 9.1; severe, 44.2 ± 9.0). Tukey honestly significant difference follow-up tests found that the difference between each adjacent level of severity (e.g., nondepressed vs minimally depressed; mild vs moderate) was significant. We repeated this analysis, limiting the sample to the 694 patients diagnosed with MDD and bipolar depression. This truncated the range of CGI ratings, and all but 2 patients were rated mild, moderate, or severe. Despite the reduced variability in CGI scores, the overall analysis of variance was still significant (F = 45.5; df = 2,691; $P < .001$), as were all 3 follow-up comparisons (mild, 32.9 ± 8.7; moderate, 38.3 ± 8.7; severe, 44.2 ± 8.1).

4.4. Association with psychiatric diagnosis

Patients with current *DSM-IV* MDD or bipolar depression scored higher than patients not in an episode of major depression (39.2 ± 9.3 vs 22.8 ± 11.9; df = 1473; t = 29.6; $P < .001$). The diagnostic performance of the CUDOS was evaluated by comparing it to the results of the SCID interview. Fig. 2 shows the ROC curve. The area under the curve (.86) was significant ($P < .001$). The diagnostic performance of the CUDOS was evaluated by scoring it according to the *DSM-IV* diagnostic algorithm. A rating of 3 or 4 indicated symptom presence. The sensitivity of the scale was 83.3%; specificity, 72.1%; positive predictive value,

Table 4
Discriminant and convergent validity of the CUDOS symptom subscale

Scale	Correlation with CUDOS (r)
Measures of depression	
Beck Depression Inventory	.81
Extracted Hamilton Rating Scale for Depression	.69
Clinical Global Index of Depression Severity	.71
Measures of nondepressive symptom domains	
Eating Disorder Inventory–bulimia subscale	.35
Eating Disorder Inventory–anorexia subscale	.22
Self Report Mania Inventory	.36
Brief Fear of Negative Evaluation Scale	.41
Fear Questionnaire–social phobia subscale	.37
Fear Questionnaire–agoraphobia subscale	.35
Social Phobia and Agoraphobia Inventory–agoraphobia subscale	.44
Symptom Rating Test–paranoia subscale	.44
Symptom Rating Test–psychosis subscale	.25
Posttraumatic Stress Disorder Scale	.39
Maudsley Obsession-Compulsion Questionnaire	.48
Penn State Worry Scale	.46
Beck Anxiety Inventory	.53
Michigan Alcohol Screening Test	.08
Drug Abuse Screening Test	.12
Whitely Index	.27
Somatic Symptom Index	.44

All correlations are significant at $P < .001$ except Michigan Alcohol Screening Test (nonsignificant) and Drug Abuse Screening Test ($P < .05$).

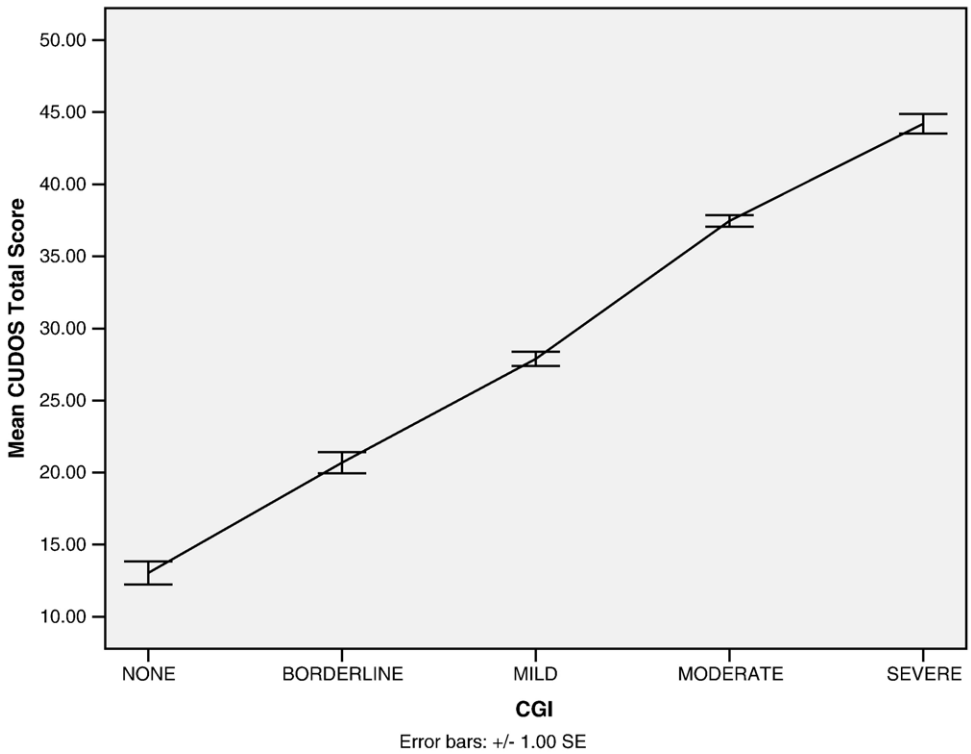

Fig. 1. Levels of depression severity and mean CUDOS total scores in 1475 psychiatric patients.

72.6%; negative predictive value, 82.9%; and chance-corrected agreement level, $k = .55$.

4.5. Sensitivity to change

Fifty-five depressed patients completed the CUDOS a second time 6 weeks after initiating antidepressant medication. At follow-up, CUDOS scores significantly declined (38.2 ± 9.0 vs 19.8 ± 12.3, paired $t = 9.63$, $P < .001$). Likewise, scores on the MADRS were significantly lower at follow-up (28.7 ± 5.8 vs 11.3 ± 9.0, paired $t = 13.27$, $P < .001$). On the CGI, 22 patients were rated very much improved, 17 patients were rate much improved, and 16 patients were rated minimally or unimproved. The CUDOS scores after 6 weeks of treatment were lowest in the patients that had the most robust improvement (very much improved, 9.4 ± 6.7; much improved, 20.7 ± 8.2; minimal/no improvement, 33.2 ± 7.9; F = 46.5; $df = 2,52$; $P < .001$). Tukey's HSD follow-up tests found that the difference between each adjacent level of improvement (e.g., very much improved vs much improved; much improved vs minimal/no improvement) was significant. We divided the patients into 3 groups using the MADRS to determine remission and response status. Patients scoring 10 and below at week 6 were considered to be in remission (n = 29); patients improving at least 50% on the MADRS but who scored above 10 at week 6 were considered responders (n = 7), and the remaining patients were considered nonresponders (n = 19). Consistent with the foregoing analysis based on the CGI, the overall 3-group comparison was significant (F = 34.4; $df = 2,52$; $P < .001$), and Tukey's HSD follow-up tests indicated that remitters scored significantly lower than

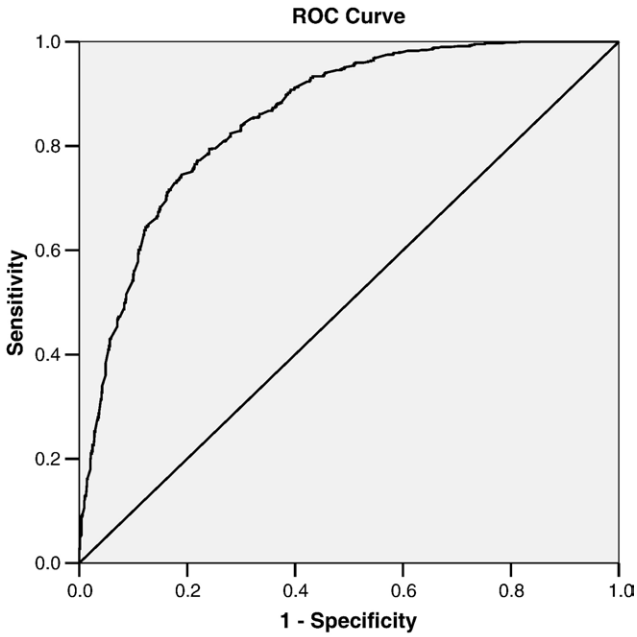

Fig. 2. Receiver operating curve for the CUDOS in detecting major depression in 1475 psychiatric outpatients.

responders (11.6 ± 8.6 vs 20.8 ± 4.2); responders scored significantly lower than nonresponders (20.8 ± 4.2 vs 31.8 ± 8.7). Based on the MADRS, the effect size of treatment was 2.28, indicating a large effect size. Based on the CUDOS the effect size was 1.70, also indicating a large effect size.

5. Discussion

The results of this large validation study of the CUDOS shows that it is a reliable and valid measure of depression. A previous study demonstrated that the scale is feasible to incorporate into routine clinical practice [15]. On average, the scale takes less than 2 minutes to complete, and more than 95% of patients were able to complete it in less than 3 minutes [15]. Patients did not find scale completion burdensome and were willing to complete it on a regular basis [15]. The CUDOS achieved high levels of internal consistency and test-retest reliability and was more highly correlated with another self-report measure of depression than with measures of anxiety, substance use problems, eating disorders, and somatization, thereby supporting the convergent and discriminant validity of the scale. The CUDOS was also significantly correlated with interviewer ratings of the severity of depression, and CUDOS scores were significantly different in depressed patients with mild, moderate, and severe levels of depression. The CUDOS was a valid measure of symptom change, and distinguished patients whose episode had remitted, responded or had not responded to treatment. And finally, the diagnostic properties of the CUDOS were adequate.

There is no shortage of self-report questionnaires that assess depression; therefore, the development of any new scale should be questioned. The CUDOS distinguishes itself from existing instruments in several respects. The CUDOS was intended as both a case-finding measure as well as a measure of depression severity that is sensitive to change. The CUDOS is one of a limited number of case-finding scales that are directly tied to the *DSM-IV* diagnostic criteria for MDD. Previously the senior author developed the Inventory to Diagnose Depression [9-11] and the Diagnostic Inventory for Depression [14] as case-finding measures for *DSM-III* and *DSM-IV* MDD, respectively. On both of these scales, the symptoms were assessed by groups of 5 statements arranged in order of increasing severity. Altogether, the scales consist of approximately 100 statements and take about 15 to 20 minutes to complete. Thus, the CUDOS is much briefer and takes less time to complete, yet its diagnostic properties are nearly as good as the longer scales. Scale brevity may be less of an issue when the measure is used a single time for screening/diagnostic purposes; however, when also used as a repeated measure to monitor the course of treatment, then scale length assumes greater importance. Although self-administered questionnaires are not impositions on the clinicians' time, they are nonetheless a burden on patients' time, and it can be difficult to use a relatively long questionnaire at every visit. Despite its brevity, we found that the CUDOS correlated highly with longer questionnaire assessments of depression and correlated highly with clinician evaluations of depression severity.

Determination of whether someone is a case on the CUDOS follows the *DSM-IV* algorithmic approach. This is in contrast to most other depression scales in which total scale scores are computed, and caseness is based on whether the individual scores above or below a threshold score. A vexing problem with this latter approach is that different studies find that different thresholds are optimal for distinguishing cases from noncases. For example, studies of the BDI as a case-finding instrument have used cutoff points of 10 [43-46], 11 [47,48], 12 [49] 13 [50,51], 15 [52], 16 [53], 23 [54] and even as high as 29 [55] to classify individuals as depressed or not. Admittedly, the BDI was not intended by its developers to be a diagnostic or screening instrument. However, even when a scale's developers recommend a particular threshold to identify cases based on a total scale score, researchers will use varying cutoff scores. For example, the Center for Epidemiological Studies–Depression [13] scale was developed as a brief screening tool with a recommended threshold score to identify depressed individuals. Other investigators, however, have used different cutoff scores to identify depressed cases [56].

Our data and clinical experience allowed us to approximate ranges of scores corresponding to a dimensional assessment of depression severity. We recommend that the nondepressed range corresponds to CUDOS scores of 0 to10; minimal depression, 11 to 20; mild depression, 21 to 30; moderate depression, 31 to 45; and severe depression, 46 and above.

The 9-item Patient Health Questionnaire (PHQ-9) [57] is another brief self-report measure assessing each of the 9 *DSM-IV* diagnostic criteria of MDD using a Likert scale similar to the CUDOS. In fact, because it contains fewer items than the CUDOS, it probably takes even less time to complete. However, the advantage offered by being somewhat briefer is offset by some loss of information. The PHQ-9 adheres to the construction of the *DSM-IV* criteria; thus, compound *DSM-IV* criteria, which refer to more than one symptom (eg, insomnia or hypersomnia; increased or decreased appetite), are represented by a single item on PHQ-9. Since treatment decision-making might be influenced by whether a patient has insomnia or is sleeping too much, or has a reduced appetite or is eating too much, the PHQ-9 does not capture potentially clinically significant information.

Elsewhere, we reported that patients did not perceive the CUDOS as burdensome to complete and were willing to complete it routinely in order to assist their clinician in monitoring their progress [15]. A consumer-friendly reliable and valid self-administered questionnaire can improve the efficiency of the clinical encounter and allow clinicians to

spend more time evaluating and discussing topics other than symptoms. In this era when many clinical encounters are 15-minute medication visits, increased efficiency can make the visit more meaningful and beneficial to both clinicians and patients. The brevity of the CUDOS lends itself to regular administration in clinical practice. Although brief, it nonetheless covers the full range of *DSM-IV* diagnostic criteria and thus provides clinically useful information.

Another advantage of the CUDOS, resulting from its direct tie to the *DSM-IV* criteria, is that it is possible to follow the *DSM-IV* guidelines when evaluating episode remission. According to *DSM-IV*, episode remission is based in large part on the number of residual diagnostic criteria that are present at the time of the evaluation. Thus, the CUDOS could be used to evaluate the completeness of episode remission (ie, partial vs full remission). The most frequent approach towards defining remission in treatment efficacy studies is based on scoring below a threshold on a clinician rated symptom severity measure such as the Hamilton Depression Scale or Montgomery-Asberg Depression Rating Scale. Previously, we determined the cutoff of the CUDOS that corresponded to the most widely used cutoff on the Hamilton to define remission [58] and validated the CUDOS' ability to distinguish between remitters and nonremitters amongst treatment responders [59]. In the present study, we found that scores on the CUDOS were significantly different in patients rated much or very much improved on the CGI. In addition, CUDOS scores were significantly different in treatment responders and remitters defined according to the MADRS.

Finally, the CUDOS is unique in that it is the only self-administered depression scale that not only evaluates the symptoms of depression but also assesses both psychosocial impairment due to depression and quality of life. We described the reliability and validity of these single item questions of these domains elsewhere [7]. The importance of these constructs has been increasingly recognized during the past decade; however, no self-report scale of depressive symptoms evaluates all of these domains.

In conclusion, the CUDOS is a reliable and valid brief self-administered depression questionnaire that is tied to the *DSM-IV* criteria and can be incorporated into routine clinical practice without significant intrusion on patients', clinicians', or support staffs' time. Although the results of this large validation study are encouraging, they require replication in samples with different demographic and clinical characteristics.

Appendix A

INSTRUCTIONS

This questionnaire includes questions about symptoms of depression. For each item please indicate how well it describes you during the PAST WEEK, INCLUDING TODAY. Circle the number in the columns next to the item that best describes you.

RATING GUIDELINES
0=not at all true (0 days)
1=rarely true (1-2 days)
2=sometimes true (3-4 days)
3=often true (5-6 days)
4=almost always true (every day)

During the PAST WEEK, INCLUDING TODAY....					
1. I felt sad or depressed	0	1	2	3	4
2. I was not as interested in my usual activities	0	1	2	3	4
3. My appetite was poor and I didn't feel like eating	0	1	2	3	4
4. My appetite was much greater than usual	0	1	2	3	4
5. I had difficulty sleeping	0	1	2	3	4
6. I was sleeping too much	0	1	2	3	4
7. I felt very fidgety, making it difficult to sit still	0	1	2	3	4
8. I felt physically slowed down, like my body was stuck in mud	0	1	2	3	4
9. My energy level was low	0	1	2	3	4
10. I felt guilty	0	1	2	3	4
11. I thought I was a failure	0	1	2	3	4
12. I had problems concentrating	0	1	2	3	4
13. I had more difficulties making decisions than usual	0	1	2	3	4
14. I wished I was dead	0	1	2	3	4
15. I thought about killing myself	0	1	2	3	4
16. I thought that the future looked hopeless	0	1	2	3	4

17. Overall, how much have symptoms of depression interfered with or caused difficulties in your life during the past week?
    0) not at all
    1) a little bit
    2) a moderate amount
    3) quite a bit
    4) extremely
18. How would you rate your overall quality of life during the past week?
    0) very good, my life could hardly be better
    1) pretty good, most things are going well
    2) the good and bad parts are about equal
    3) pretty bad, most things are going poorly
    4) very bad, my life could hardly be worse

References

[1] Gilbody S, House A, Sheldon T. Psychiatrists in the UK do not use outcomes measures. Br J Psychiatry 2002;180:101-3.
[2] Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56-62.
[3] Zimmerman M, Mattia JI. The Psychiatric Diagnostic Screening Questionnaire: Development, reliability and validity. Compr Psychiatry 2001;42:175-89.
[4] Zimmerman M, Mattia JI. A self-report scale to help make psychiatric diagnoses: The Psychiatric Diagnostic Screening Questionnaire (PDSQ). Arch Gen Psychiatry 2001;58:787-94.
[5] Zimmerman M, Chelminski I. A scale to screen for axis I disorders in psychiatric outpatients: Performance of the Psychiatric Diagnostic Screening Questionnaire. Psychol Med 2006;36:1601-11.
[6] Zimmerman M, Posternak M, Chelminski I. A Standardized Clinical Outcome Rating Scale for Depression for use in clinical practice. Depress Anxiety 2005;22:36-40.
[7] Zimmerman M, Ruggero C, Chelminski I, Young D, Posternak M, Friedman M, et al. Developing brief measures for use in clinical

practice: The reliability and validity of single-item self-report measures of depression symptom severity, psychosocial impairment due to depression and quality of life. J Clin Psychiatry 2006;67:1536-41.
[8] Nezu A, Ronan G, Meadows E, McClure K. Practitioner's guide to empirically based measures of depression. Kluwer Academic/Plenum Publishers: New York; 2000.
[9] Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561-71.
[10] Zimmerman M, Coryell W, Corenthal C, Wilson S. A self-report scale to diagnose major depression disorder. Arch Gen Psychiatry 1986;43: 1076-81.
[11] Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The Inventory of Depressive Symptomatology (IDS). Psychol Med 1996; 26:477-86.
[12] Zung WWK. A self-rating depression scale. Arch Gen Psychiatry 1965;12:63-70.
[13] Radloff LS. The CES-D scale: A self-report depression scale for research in the general population. Appl Psychol Meas 1977;3: 385-401.
[14] Zimmerman M, Sheeran T, Young D. The Diagnostic Inventory for Depression: A self-report scale to diagnose *DSM-IV* for major depressive disorder. J Clin Psychol 2004;60:87-110.
[15] Zimmerman M, McGlinchey J. Depressed patients' acceptability of the use of self-administered scales to measure outcome in clinical practice. [submitted].
[16] First MB, Spitzer RL, Williams JBW, Gibbon M. Structured Clinical Interview for *DSM-IV* (SCID). Washington, D.C.: American Psychiatric Association; 1997.
[17] Endicott J, Spitzer RL. A diagnostic interview: the schedule for affective disorders and schizophrenia. Arch Gen Psychiatry 1978;35: 837-44.
[18] Endicott J, Cohen J, Nee J, Fleiss J, Sarantakos S. Hamilton depression rating scale. Arch Gen Psychiatry 1981;38:98-103.
[19] Guy W. ECDEU assessment manual for psychopharmacology. Rockville, MD: National Institute of Mental Health; 1976.
[20] Zimmerman M, Mattia JI. Psychiatric diagnosis in clinical practice: Is comorbidity being missed? Compr Psychiatry 1999; 40:182-91.
[21] Zimmerman M, Mattia JI. Differences between clinical and research practice in diagnosing borderline personality disorder. Am J Psychiatry 1999;156:1570-4.
[22] Zimmerman M. Integrating the assessment methods of researchers in routine clinical practice: The Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project. In: First M, editor. Standardized Evaluation in Clinical Practice, vol. 22. Washington, DC: American Psychiatric Publishing, Inc; 2003. p. 29-74.
[23] Garner DM, Olmstead MP, Polivy J. Development and validation of a multidimensional eating disorder inventory for anorexia nervosa and bulimia. Int J Eat Disord 1983;2:15-34.
[24] Beck AT, Rush AJ, Shaw BF, Emery G. Cognitive therapy of depression. New York, NY: The Guilford Press; 1979.
[25] Leary MR. A brief version of the Fear of Negative Evaluation Scale. Pers Soc Psychol Bull 1983;9:371-5.
[26] Marks IM, Mathews AM. Brief standard self-rating for phobic patients. Behav Res Ther 1979;17:263-7.
[27] Turner SM, Beidel DC, Dancu CV, Stanley MA. An empirically derived inventory to measure social fears and anxiety: The Social Phobia and Anxiety Inventory. Psychol Assess: Journal of Consulting and Clinical Psychology 1989;1:35-40.
[28] Foa EB, Riggs DS, Dancu CV, Rothbaum BO. Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. J Trauma Stress 1993;6:459-73.
[29] Hodgson RJ, Rachman S. Obsessional compulsive complaints. Behav Res Ther 1977;10:111-7.
[30] Meyer TJ, Miller ML, Metzger RL, Borkovec TD. Development and validation of the Penn State Worry Questionnaire. Behav Res Ther 1990;28:487-95.
[31] Beck AT, Brown G, Epstein N, Steer R. An inventory for measuring clinical anxiety: Psychometric properties. J Consult Clin Psychol 1988; 56:893-7.
[32] Selzer ML. The Michigan Alcoholism Screening Test: The quest for a new diagnostic instrument. Am J Psychiatry 1971;127:1653-8.
[33] Skinner HA. The Drug Abuse Screening Test. Addict Behav 1982;7: 363-71.
[34] Pilowsky I. Dimensions of hypochondriasis. Br J Psychiatry 1967;113: 89-93.
[35] Othmer E, DeSouza C. A screening test for somatization disorder (hysteria). Am J Psychiatry 1985;142:1146-9.
[36] Swartz M, Hughes D, George L, Blazer D, Landerman R, Bucholz K. Developing a screening index for community studies of somatization disorder. J Psychiatr Res 1986;20:335-43.
[37] Shugar G, Schertzer S, Toner BB, DiGasbarro I. Development, use, and factor analysis of a self-report inventory for mania. Compr Psychiatry 1992;33:325-31.
[38] Kellner R. The Symptom-Rating Test. In: Sartorius N, Ban TA, editors. Assessment of Depression. New York: Springer-Verlag; 1985.
[39] Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979;134:382-9.
[40] Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait multi-method matrix. Psychol Bull 1959; 56:81-105.
[41] Hsiao JK, Bartko JJ, Potter WZ. Diagnosing diagnoses: Receiver operating characteristic methods and psychiatry. Arch Gen Psychiatry 1989;46:664-7.
[42] Mossman D, Somoza E. Maximizing diagnostic information from the dexamethasone suppression test. Arch Gen Psychiatry 1989;46:653-60.
[43] Craven JL, Rodin GM, Littlefield C. The Beck Depression Inventory as a screening device for major depression in renal dialysis patients. Int J Psychiatry Med 1988;18:365-74.
[44] Deardorff WW, Funabiki D. A diagnostic caution in screening for depressed college students. Cogn Ther Res 1985;9:277-84.
[45] Oliver JM, Simmons ME. Depression as measured by the *DSM-III* and the Beck Depression Inventory in an unselected adult population. J Consult Clin Psychol 1984;52:892-8.
[46] Zich JM, Attkisson CC, Greenfield TK. Screening for depression in primary care clinics: The CES-D and the BDI. Int J Psychiatry Med 1990;20:259-77.
[47] Gallagher D, Breckenridge J, Steinmetz J, Thompson L. The Beck Depression Inventory and Research Diagnostic Criteria: Congruence in an older population. J Consult Clin Psychol 1983;51:945-6.
[48] Harris B, Huckle P, Thomas R, Hohns S, Fung H. The use of rating scales to identify post-natal depression. Br J Psychiatry 1989;154: 813-7.
[49] Lasa L, Ayuso-Mateos JL, Vazquez-Barquero JL, Diez-Manrique FJ, Dowrick CF. The use of the Beck Depression Inventory to screen for depression in the general population: a preliminary analysis. J Affect Disord 2000;57:261-5.
[50] Hesselbrock MN, Hesselbrock VM, Tennen H, Meyer RE, Workman KL. Methodological considerations in the assessment of depression in alcoholics. J Consult Clin Psychol 1983;51:399-405.
[51] Turner JA, Romano JM. Self-reporting screening measures for depression in chronic pain patients. J Clin Psychol 1984;40:909-13.
[52] Bishop SR, Edgley K, Fisher R, Sullivan MJL. Screening for depression in chronic low back pain with the Beck Depression Inventory. Can J Rehabil 1993;7:143-8.
[53] Holcomb WL, Stone LS, Lustman PJ, Gavard JA, Mostello DJ. Screening for depression in pregnancy: Characteristics of the Beck Depression Inventory. Obstet Gynecol 1996;88:1021-5.
[54] Martinsen EW, Friis S, Hoffart A. Assessment of depression: Comparison between Beck Depression Inventory and subscales of Comprehensive Psychopathological Rating Scale. Acta Psychiatr Scand 1995;92:460-3.
[55] Lykouras L, Oulis P, Adrachta D, Daskalopoulou E, Kalfakis N, Triantaphyllou N, et al. Beck Depression Inventory in the detection of

depression among neurological inpatients. Psychopathology 1998;31: 213-9.

[56] Coulehan JL, Schulberg HC, Block MR, Janosky JE, Arena VC. Depressive symptomatology and medical co-morbidity in a primary care clinic. Int J Psychiatry Med 1990;20:335-47.

[57] Kroenke K, Spitzer R, Williams J. The PHQ-9. Validity of a brief depression severity measure. J Gen Intern Med 2001;16:606-13.

[58] Zimmerman M, Posternak M, Chelminski I. Using a self-report depression scale to identify remission in depressed outpatients. Am J Psychiatry 2004;161:1911-3.

[59] Zimmerman M, Posternak M, McGlinchey J, Friedman M, Attiullah N, Boerescu D. Validity of a self-report depression symptom scale for identifying remission in depressed outpatients. Compr Psychiatry 2006;47:185-8.