The PHQ-9

Validity of a Brief Depression Severity Measure

Kurt Kroenke, MD, Robert L. Spitzer, MD, Janet B. W. Williams, DSW

OBJECTIVE: **While considerable attention has focused on improving the detection of depression, assessment of severity is also important in guiding treatment decisions. Therefore, we examined the validity of a brief, new measure of depression severity.**

MEASUREMENTS: **The Patient Health Questionnaire (PHQ) is a self-administered version of the PRIME-MD diagnostic instrument for common mental disorders. The PHQ-9 is the depression module, which scores each of the 9 DSM-IV criteria as "0" (not at all) to "3" (nearly every day). The PHQ-9 was completed by 6,000 patients in 8 primary care clinics and 7 obstetrics-gynecology clinics. Construct validity was assessed using the 20-item Short-Form General Health Survey, self-reported sick days and clinic visits, and symptom-related difficulty. Criterion validity was assessed against an independent structured mental health professional (MHP) interview in a sample of 580 patients.**

RESULTS: **As PHQ-9 depression severity increased, there was a substantial decrease in functional status on all 6 SF-20 subscales. Also, symptom-related difficulty, sick days, and health care utilization increased. Using the MHP reinterview as the criterion standard, a PHQ-9 score $\geq$10 had a sensitivity of 88% and a specificity of 88% for major depression. PHQ-9 scores of 5, 10, 15, and 20 represented mild, moderate, moderately severe, and severe depression, respectively. Results were similar in the primary care and obstetrics-gynecology samples.**

CONCLUSION: **In addition to making criteria-based diagnoses of depressive disorders, the PHQ-9 is also a reliable and valid measure of depression severity. These characteristics plus its brevity make the PHQ-9 a useful clinical and research tool.**

KEY WORDS: **depression; diagnosis; screening; psychological tests; health status.**
J GEN INTERN MED 2001;16:606–613.

Depression is one of the most prevalent and treatable mental disorders and is regularly seen by a wide spectrum of health care providers, including mental health specialists, medical and surgical subspecialists, and primary care clinicians. There are a number of case-finding instruments for detecting depression in primary care, ranging from 2 to 28 items in length.[1,2] Typically, these can be scored as continuous measures of depression severity and also have established cut points above which the probability of major depression is substantially increased. Scores on these various measures tend to be highly correlated,[3] and it is not evident that any one measure is superior to the others.[1,2,4]

The Patient Health Questionnaire (PHQ) is a new instrument for making criteria-based diagnoses of depressive and other mental disorders commonly encountered in primary care. The diagnostic validity of the PHQ has recently been established in 2 studies involving 3,000 patients in 8 primary care clinics and 3,000 patients in 7 obstetrics-gynecology clinics.[5,6] At 9 items, the PHQ depression scale (which we call the PHQ-9) is half the length of many other depression measures, has comparable sensitivity and specificity, and consists of the actual 9 criteria upon which the diagnosis of DSM-IV depressive disorders is based. The latter feature distinguishes the PHQ-9 from other "2-step" depression measures for which, when scores are high, additional questions must be asked to establish DSM-IV depressive diagnoses. The PHQ-9 has the potential of being a dual-purpose instrument that, with the same 9 items, can establish depressive disorder diagnoses as well as grade depressive symptom severity. In this paper, we analyze data regarding the PHQ-9 to address 3 major questions:

1. What is the reliability and efficiency of the PHQ-9 in clinical practice?
2. What are the operating characteristics (sensitivity and specificity) of the PHQ-9 as a diagnostic instrument for depressive disorders?
3. What is the construct validity of the PHQ-9 as a depression severity measure in relation to functional status, disability days, and health care utilization?

METHODS

Description of the PHQ and PHQ-9

The Patient Health Questionnaire (PHQ) is a 3-page questionnaire that can be entirely self-administered by the patient.[5] The clinician scans the completed questionnaire, verifies positive responses, and applies diagnostic algorithms that are abbreviated at the bottom of each page. The PHQ assesses 8 diagnoses, divided into threshold disorders (disorders that correspond to specific DSM-IV diagnoses: major depressive disorder, panic disorder, other anxiety disorder, and bulimia nervosa), and subthreshold disorders (disorders whose criteria encompass fewer symptoms than are required for any specific DSM-IV diagnoses: other depressive disorder, probable alcohol abuse/dependence, somatoform, and binge eating disorder).

Received from the Regenstrief Institute for Health Care and Department of Medicine, Indiana University (KK), Indianapolis, Ind; and the New York State Psychiatric Institute and Department of Psychiatry, Columbia University (RLS, JBWW), New York, NY.

Address correspondence and reprint requests to Dr. Kroenke: Regenstrief Institute for Health Care, RG-6, 1050 Wishard Blvd., Indianapolis, IN 46202 (e-mail: kkroenke@regenstrief.org).

The PHQ-9 (Appendix) is the 9-item depression module from the full PHQ. Major depression is diagnosed if 5 or more of the 9 depressive symptom criteria have been present at least "more than half the days" in the past 2 weeks, and 1 of the symptoms is depressed mood or anhedonia. Other depression is diagnosed if 2, 3, or 4 depressive symptoms have been present at least "more than half the days" in the past 2 weeks, and 1 of the symptoms is depressed mood or anhedonia. One of the 9 symptom criteria ("thoughts that you would be better off dead or of hurting yourself in some way") counts if present at all, regardless of duration. As with the original PRIME-MD, before making a final diagnosis, the clinician is expected to rule out physical causes of depression, normal bereavement, and history of a manic episode.

As a severity measure, the PHQ-9 score can range from 0 to 27, since each of the 9 items can be scored from 0 (not at all) to 3 (nearly every day). An item was also added to the end of the diagnostic portion of the PHQ-9 asking patients who checked off any problems on the questionnaire: "How *difficult* have these problems made it for you to do your work, take care of things at home, or get along with other people?"

PHQ Study Samples and Procedures

From May 1997 to November 1998, 3,890 patients, 18 years or older, were invited to participate in the PHQ Primary Care Study.[5] There were 190 who declined to participate, 266 who started but did not complete the questionnaire (often because there was inadequate time before seeing their physician), and 434 whose questionnaires were not entered into the data set because the equivalent of approximately 1 page (20 items) was not completed. This resulted in the 3,000 primary care patients reported here (1,422 from 5 general internal medicine clinics and 1,578 from 3 family practice clinics). From May 1997 to March 1999, 3,636 patients, 18 years or older, were approached to participate in the PHQ Obstetrics-Gynecology (Ob-Gyn) Study.[6] There were 245 patients who declined to participate, 127 who started but did not complete the questionnaire, and 264 whose questionnaires were not entered into the data set because the equivalent of approximately 1 page was not completed. This resulted in the 3,000 subjects from 7 obstetrics-gynecology (ob-gyn) sites. All sites used one of 2 subject selection methods to minimize sampling bias: either consecutive patients for a given clinic session or every *n*th patient until the intended quota for that session was achieved. Patient characteristics are summarized in Table 1. Besides being entirely women, the ob-gyn sample had a younger average age, more Hispanic subjects, lower average education, and less medical comorbidity.

A total of 62 physicians participated in the PHQ Primary Care Study (21 general internal medicine and 41 family practice [19 of who were family practice residents]). Their mean age was 37 years (standard deviation [SD], 6.5), and 63% were male. A total of 40 physicians and 21 nurse practitioners participated in the PHQ Ob-Gyn. Their mean age was 39 years (SD, 8.9), and 48% were male.

Table 1. Characteristics of Patients in the PHQ Primary Care and Obstetrics-gynecology Studies

Patient Characteristic	Study 1 PHQ Primary Care	Study 2 PHQ Ob-gyn
Subjects, *N*	3,000	3,000
Established clinic patient, %	80	71
Mean age, y ±SD	46 ± 17	31 ± 11
Women, %	66	100
Race, %		
White	79	39
African American	13	15
Hispanic	4	39
Marital status, %		
Married	48	52
Never married	23	33
Divorced/separated/widowed	29	15
Education, %		
College graduate	27	16
Partial college	27	25
High school graduate only	33	32
Less than high school	13	27
Medical conditions, %		
Hypertension	25	2
Arthritis	11	1
Diabetes	8	1
Pulmonary	7	2

Before seeing the physician, all patients completed the PHQ. Additionally, they completed the Medical Outcomes Study Short-Form General Health Survey (SF-20).[7] The SF-20 measures functional status in 6 domains (all scores from 0 to 100; 100=best health). Also, patients estimated the number of physician visits and disability days during the past 3 months.

Mental Health Professional (MHP) Validation Interviews

To determine the agreement of PHQ diagnoses with those of MHPs, midway through the PHQ Primary Care Study, a MHP (a PhD clinical psychologist or 1 of 3 senior psychiatric social workers) attempted to interview by telephone all subsequently entered subjects who had a telephone, agreed to be interviewed, and could be contacted within 48 hours. All except 1 site participated in these validation interviews. The MHP was blinded to the results of the PHQ. The rationale and further details of the MHP telephone interview, which used the overview from the SCID[8] and diagnostic questions from the PRIME-MD, are described in the original PRIME-MD report.[9] To examine test-retest reliability, the MHP graded the 9 PRIME-MD questions assessing DSM-IV symptoms using the same 4 response options as the PHQ-9 (i.e., not at all, several days, more than half the days, nearly every day).

The 580 subjects who had a MHP interview within 48 hours of completing the PHQ were, within each site, similar

to patients not reinterviewed in terms of demographic profile, functional status, and frequency of psychiatric diagnoses. Agreement between the PHQ diagnoses and the MHP diagnoses was examined. One modification from the original PRIME-MD algorithm was necessary. The number of criteria required for diagnosing major depressive disorder could remain the same as in DSM-IV, i.e., 5 of 9 during the past 2 weeks. However, because the PHQ response set was expanded from the simple "yes/no" in the original PRIME-MD to 4 frequency levels, lowering the PHQ threshold from "nearly every day" to "more than half the days" raised the sensitivity from 37% to 73% while maintaining high specificity (94%).

Analysis

For most analyses, the PHQ-9 score was divided into the following categories of increasing severity: 0–4, 5–9, 10–14, 15–19, and 20 or greater. These categories were chosen for several reasons. The first was pragmatic, in that the cut points of 5, 10, 15, and 20 are simple for clinicians to remember and apply. The second reason was empiric, in that using different cut points did not noticeably change the associations between increasing PHQ-9 severity and measures of construct validity.

For analyses assessing the operating characteristics of various PHQ-9 intervals or cut points, diagnostic status (major depressive disorder, other depressive disorder, or no depressive disorder) was that assigned by the independent MHP structured psychiatric interview. The latter is considered the criterion standard and provides the most conservative estimate of the operating characteristics of the PHQ-9 score. Besides calculating sensitivity and specificity of the PHQ-9 over various intervals, we also determined likelihood ratios[10] and conducted ROC curve analysis[11] as quantitative methods for combining sensitivity and specificity into a single metric.

Construct validity of the PHQ-9 as a measure of depression severity was assessed by examining functional status (the 6 SF-20 scales), disability days, symptom-related difficulty, and health care utilization (clinic visits) over the 5 PHQ-9 intervals. Analysis of covariance was used, with PHQ-9 category as the independent variable and adjusting for age, gender, race, education, study site, and number of physical disorders. Bonferroni's correction was used to adjust for multiple comparisons.

RESULTS

Reliability and Efficiency of the PHQ-9

The internal reliability of the PHQ-9 was excellent, with a Cronbach's α of 0.89 in the PHQ Primary Care Study and 0.86 in the PHQ Ob-Gyn Study. Test-retest reliability of the PHQ-9 was also excellent. Correlation between the PHQ-9 completed by the patient in the clinic and that administered telephonically by the MHP within 48 hours was 0.84, and the mean scores were nearly identical (5.08 vs 5.03).

In 85% of cases clinicians required less than 3 minutes to review responses on the full 3-page PHQ,[5] which consists of 5 modules and 28 to 58 items (depending upon the number of skip-outs). Although time to review the PHQ depression items was not measured separately, it is unlikely this took more than a minute, since the PHQ-9 includes less than one third of the items contained in the full PHQ.

Distribution of PHQ-9 Scores According to Depression Diagnostic Status

Table 2 shows the distribution of PHQ-9 scores according to depression diagnostic status in the 580 patients interviewed by a mental health professional who was blinded to the PHQ-9 results. The mean PHQ-9 score was 17.1 (SD, 6.1) in the 41 patients diagnosed by the MHP as having major depression; 10.4 (SD, 5.4) in the 65 patients diagnosed as other depressive disorder; and 3.3 (SD, 3.8) in the 474 patients with no depressive disorder. The vast majority of patients (93%) with no depressive disorder had a PHQ-9 score less than 10, while most patients (88%) with major depression had scores of 10 or greater. Scores less than 5 almost always signified the absence of a depressive disorder; scores of 5 to 9 predominantly represented patients with either no depression or subthreshold (i.e., other) depression; scores of 10 to 14 represented a spectrum of patients; and scores of 15 or greater usually indicated major depression.

Criterion Validity of PHQ-9 Assessed by Mental Health Professional Interview

Because PHQ-9 scores in the 10 to 15 range appear to represent an important "gray zone," we conducted a more detailed examination of the operating characteristics of various cut points in this range. Table 3 displays the sensitivity, specificity, and likelihood ratios for different PHQ-9 thresholds in diagnosing major depression in the 580 patients who had a MHP interview. For example, a patient with major depression is 6 times more likely than a

Table 2. Distribution of PHQ-9 Scores According to Depression Diagnostic Status*

Level of Depression Severity, PHQ-9 Score	Major Depressive Disorder (*N* = 41) *n* (%)	Other Depressive Disorder (*N* = 65) *n* (%)	No Depressive Disorder (*N* = 474) *n* (%)
Minimal, 0–4	1 (2.4)	8 (12.3)	348 (73.4)
Mild, 5–9	4 (9.8)	23 (35.4)	93 (19.6)
Moderate, 10–14	8 (19.5)	17 (26.1)	23 (4.9)
Moderately severe, 15–19	14 (34.1)	14 (21.5)	8 (1.7)
Severe, 20–27	14 (34.1)	3 (4.6)	2 (0.4)

** Depression diagnostic status was determined in 580 primary care patients by having a mental health professional who was blinded to the PHQ-9 score administer a structured psychiatric interview.*

Table 3. Operating Characteristics of Various PHQ-9 Cutpoints for Diagnosing Major Depression*

PHQ-9 Depression Score	Sensitivity (%)	Specificity (%)	Likelihhod Ratio
≥9	95	84	6.0
≥10	88	88	7.1
≥11	83	89	7.8
≥12	83	92	10.2
≥13	78	93	11.1
≥14	73	94	12.0
≥15	68	95	13.6

* *In 580 patients who underwent a structured psychiatric interview by a mental health professional to determine the presence or absence of major depression using DSM-IV diagnostic criteria.*

patient without major depression to have a PHQ-9 score of 9 or greater and 13.6 times more likely to have a score of 15 or greater. In this sample with a 7% prevalence of major depression (41 out of 580 patients), the positive predictive value for major depression ranged from 31% for a PHQ-9 cut point of 9 to 51% for a cut point of 15.

Examination of likelihood ratios further confirmed the substantial association between increasing PHQ-9 scores and the likelihood of major depression. The positive likelihood ratios of PHQ-9 scores of 0–4, 5–9, 10–14, 15–19, and 20–27 for major depression were 0.04, 0.5, 2.6, 8.4, and 36.8, respectively. Interpretation of these likelihood ratios means that, for example, a PHQ-9 score in the 0–4 ranges is only 0.04 (i.e., 1/25) times as likely in a patient with major depression compared to a patient without major depression, while a score of 10 to 14 is 2.6 times as likely and a score of 15 to 19 is 8.4 times as likely. The positive likelihood ratio of these same 5 PHQ-9 intervals for any depression (i.e., major or other depressive disorder) was 0.12, 1.3, 4.9, 15.7, and 38.0, respectively.

ROC analysis showed that the area under the curve for the PHQ-9 in diagnosing major depression was 0.95, suggesting a test that discriminates well between persons with and without major depression. The area under the curve for the 5-item mental health scale of the SF-20 was 0.93.

Construct Validity of PHQ-9 Assessed by Functional Status and other Measures

As shown in Table 4, there was a strong association between increasing PHQ-9 depression severity scores and worsening function on all 6 SF-20 scales. Several findings should be noted. First, results were essentially the same for both the primary care and obstetrics-gynecology samples. Second, the monotonic decrease in SF-20 scores with increasing PHQ-9 scores were greatest for the scales that previous studies have shown should be most strongly related to depression, i.e., mental health, followed by social, overall, and role functioning, with a lesser relationship to pain and physical functioning.[12] Third, most pairwise comparisons within each SF-20 scale between successive PHQ-9 levels were highly significant.

Figure 1 illustrates graphically the relationship between increasing PHQ-9 scores and worsening functional status. Decrements in SF-20 scores are shown in terms of effect size, which is the difference in mean SF-20 scores, expressed as the number of standard deviations, between each PHQ-9 interval subgroup and the reference group. The reference group is the group with the lowest PHQ-9 scores (i.e., 0–4), and the standard deviation used is that of

Table 4. Relationship Between PHQ-9 Depression Score and SF-20 Health-related Quality of Life Scales*

Level of Depression Severity, PHQ-9 Score	Mean (95% CI) SF-20 Scale Score											
	Mental		Social		Role		General		Pain		Physical	
	Primary Care	Ob-gyn	Primary Care	Ob-gyn	Primary Care	Ob-gyn	Primary Care	Ob-gyn	Primary Care	Ob-gyn	Primary Care	Ob-gyn
Minimal, 1–4	81 (80 to 82)	81 (80 to 82)	92 (91 to 93)	91 (90 to 92)	86 (84 to 88)	88 (87 to 90)	70 (69 to 71)	75 (73 to 76)	66 (65 to 68)	73 (72 to 74)	83 (81 to 83)	86 (85 to 87)
Mild, 5–9	65 (64 to 66)	66 (64 to 67)	77 (75 to 79)	81 (79 to 83)	63 (60 to 66)	77 (74 to 79)	50 (48 to 52)	57 (55 to 58)	52[a] (50 to 54)	59[a] (57 to 61)	69 (67 to 71)	76[a] (74 to 77)
Moderate, 10–14	51 (50 to 53)	53 (51 to 55)	65 (62 to 68)	75[a] (72 to 78)	53[a] (49 to 58)	64[a] (60 to 69)	40[a] (37 to 43)	48 (45 to 51)	49[a] (45 to 52)	53[a,b] (50 to 57)	63[a] (60 to 66)	74[a] (71 to 77)
Moderately severe, 15–19	43 (40 to 45)	45 (42 to 48)	55 (51 to 59)	68[a] (63 to 72)	42[a] (36 to 48)	64[a,b] (57 to 71)	33[a,b] (29 to 37)	40[a] (35 to 44)	45[a,b] (41 to 50)	50[b] (45 to 55)	57[a,b] (53 to 61)	74[a] (69 to 78)
Severe, 20–27	29 (25 to 31)	35 (31 to 39)	40 (35 to 44)	50 (43 to 56)	27 (20 to 35)	48[b] (39 to 58)	27[b] (22 to 31)	30[a] (24 to 36)	40[b] (35 to 45)	46[b] (40 to 53)	53[b] (48 to57)	56 (50 to 62)

* *SF-20 scores are adjusted for age, gender, race, education, study site, and number of physical disorders. Point estimates for the mean as well as 95% confidence intervals (±1.96 × standard error of the mean) are displayed.*
Most pairwise comparisons of mean SF-20 scores between each PHQ-9 level within each scale are significant at P < 0.05 using Bonferroni's correction for multiple comparisons. Only those pairwise comparisons that share a common superscript letter (a, b, or a,b) are not significant.

the entire sample. Effect sizes of 0.5 and 0.8 are typically considered moderate and large between-group differences, respectively.[13] Figure 1 shows effect sizes for the primary care sample; results for the obstetrics-gynecology sample (not displayed) were similar.

When the PHQ-9 was examined as a continuous variable, its strength of association with the SF-20 scales was concordant with the pattern seen in Figure 1. The PHQ-9 correlated most strongly with mental health (0.73), followed by general health perceptions (0.55), social functioning (0.52), role functioning (0.43), physical functioning (0.37), and bodily pain (0.33).

Table 5 shows the association between PHQ-9 severity levels and 3 other measures of construct validity: self-reported disability days, clinic visits, and the general amount of difficulty patients attribute to their symptoms. Greater levels of depression severity were associated with a monotonic increase in disability days, health-care utilization, and symptom-related difficulty in activities and relationships. When the PHQ-9 was examined as a continuous variable, its correlation was 0.39 with disability days, 0.24 with physician visits, and 0.55 with symptom-related difficulty.

Because our sample was relatively young and disproportionately female, we examined the influence of age and gender in several ways. First, simple correlations between PHQ-9 score and measures of construct validity were similar when examined separately for women and men, while correlations were somewhat lower but still highly significant in patients 65 years and older compared to younger individuals. Second, analysis of covariance results showed age had an independent and weak effect on only one outcome (SF-20 physical functioning), while gender had no independent effect.

The single item assessing difficulty that the patients attributed to their depressive symptoms correlated strongly with impairment as measured by the SF-20 subscales,

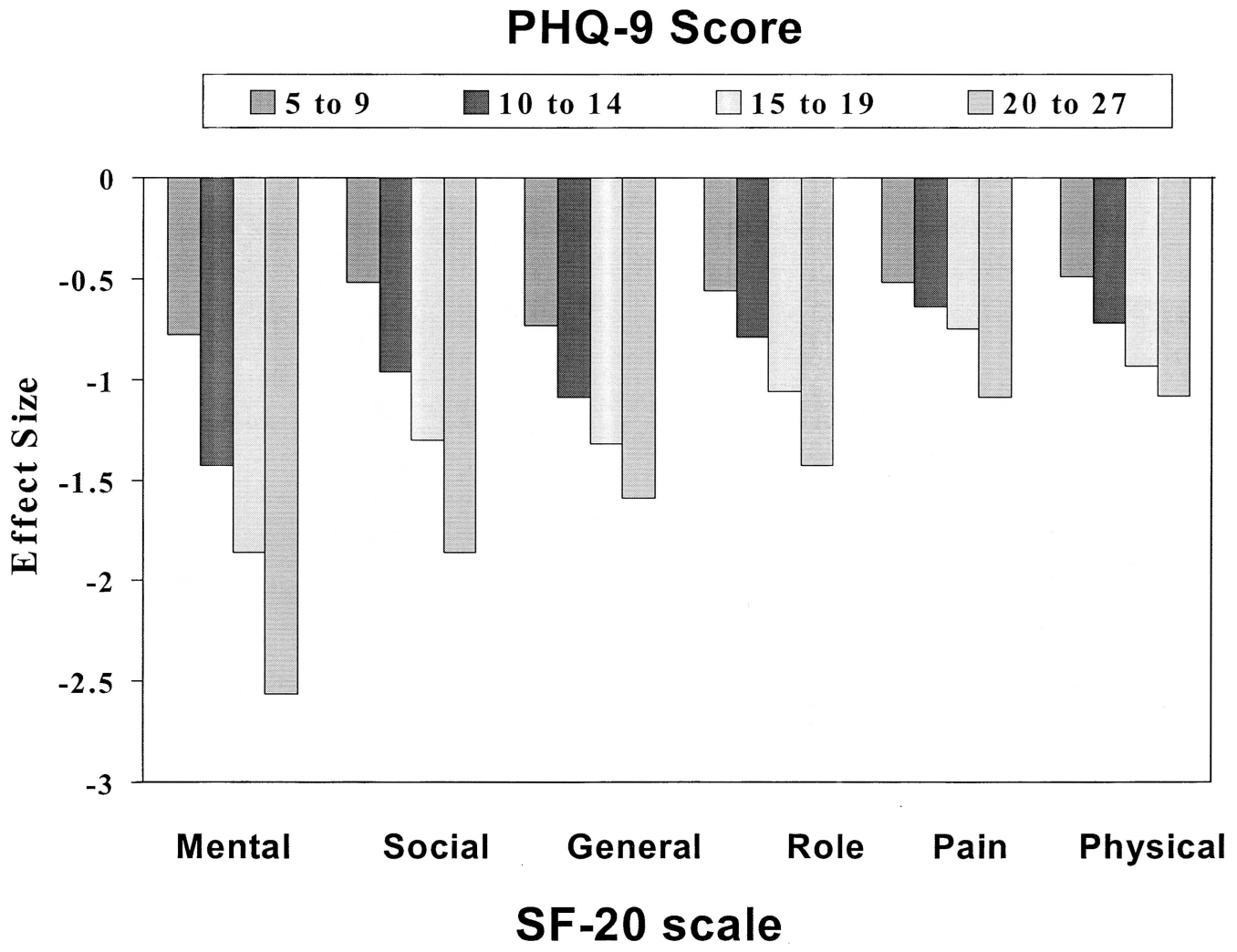

FIGURE 1. Relationship between depression severity as measured by the PHQ-9 and decline in functional status as measured by the 6 subscales of the SF-20. The decrement in SF-20 scores are shown as the difference between each PHQ-9 severity group and the nondepressed reference group (i.e., those with PHQ-9 scores of 0 to 4). Effect size is the difference in group means divided by the standard deviation of the entire sample.

Table 5. Relationship Between PHQ-9 Depression Severity Score and Disability Days, Symptom-related Difficulty, and Clinic Visits

Level of Depression Severity, PHQ-9 Score	Mean Disability Days (95% CI)*		Symptom-related Difficulty (%)†		Mean Physician Visits (95% CI)*	
	Primary Care	Obstetrics-gynecology	Primary Care	Obstetrics-gynecology	Primary Care	Obstetrics-gynecology
Minimal, 1–4	2.4 (1.7 to 3.1)	2.2(1.7 to 2.7)	1.5	0.6	1.0 (0.9 to 1.1)	0.9[a] (0.8 to 1.0)
Mild, 5–9	6.7 (5.5 to 7.8)	5.8 (4.9 to 6.6)	10.2	4.8	1.8[a] (1.6 to 2.0)	0.9[a] (1.0 to 1.4)
Moderate, 10–14	11.4 (9.5 to 13.1)	9.9[a] (8.4 to 11.3)	24.4	16.8	2.0[a] (1.7 to 2.4)	1.3[a] (1.0 to 1.6)
Moderately severe, 15–19	16.6 (14.1 to 19.0)	10.8[a] (8.6 to 13.0)	45.1[a]	36.0	2.4[a] (1.9 to 2.8)	2.3[b] (1.8 to 2.8)
Severe, 20–27	28.1 (25.2 to 31.0)	13.8[a] (10.8 to 16.7)	57.1[a]	56.6	3.7 (3.2 to 4.2)	2.3[b] (1.7 to 3.0)

** Disability days refers to number of days in past 3 months that their symptoms interfered with their usual activities. Physician visits refers to past 3 months also. Both are self-report. Means are also adjusted for age, gender, race, education, study site, and number of physical disorders.*

† Response to single question: "How difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?" The 4 response categories are "not difficult at all," "somewhat difficult," "very difficult," and "extremely difficult." Report difficulty in this table refers to those patients reporting "very" or "extremely" difficult.

Most pairwise comparisons between each PHQ-9 severity level for a given variable are significant at $P < 0.05$ using Bonferroni's correction for multiple comparisons. Only those pairwise comparisons that share a common superscript letter (a, b, or a,b) are not significant.

particularly those domains known to be most affected by mental disorders. Correlations of the single symptom-related difficulty item with the SF-20 scales in the primary care sample were 0.53 for mental health, 0.42 for general health perceptions, 0.40 for social functioning, 0.38 for role functioning, 0.27 for bodily pain, and 0.27 for physical functioning. Although slightly lower in the obstetrics-gynecology sample, correlations showed a similar rank order.

DISCUSSION

Data from our 2 studies totaling 6,000 patients provide strong evidence for the validity of the PHQ-9 as a brief measure of depression severity. Criterion validity was demonstrated in the sample of 580 primary care patients who underwent an independent reinterview by a mental health professional. Construct validity was established by the strong association between PHQ-9 scores and functional status, disability days, and symptom-related difficulty. External validity was achieved by replicating the findings from the 3,000 primary care patients in a second sample of 3,000 obstetrics-gynecology patients. Indeed, the similar results seen in rather different patient populations suggests our PHQ-9 findings may be generalizable to outpatients seen in a variety of clinic settings.

Our analysis of the full range of PHQ-9 scores complements rather than supercedes the validated PHQ-9 algorithm for establishing categorical diagnoses. However, as the PHQ-9 is increasingly used as a continuous measure of depression severity, it will be helpful to know the probability of a major or subthreshold depressive disorder at various cut points. PHQ-9 scores of 5, 10, 15, and 20 represent valid and easy-to-remember thresholds demarcating the lower limits of mild, moderate, moderately severe, and severe depression. In particular, scores less than 10 seldom occur in individuals with major depression while scores of 15 or greater usually signify the presence of major depression. In the "gray zone" of 10 to 14, increasing PHQ-9 scores are associated, as expected, with increasing specificity and declining sensitivity. However, the operating characteristics of the PHQ-9 displayed at various cut points in Table 2 compare favorably to 9 other case-finding instruments for depression in primary care which have an overall sensitivity of 84%, a specificity of 72%, and a positive likelihood ratio of 2.86.[1] Likewise, the positive predictive value of the PHQ-9 (ranging from 31% to 51% depending upon the cut point) is similar to other instruments; of note, predictive value is related not only to a measure's sensitivity and specificity but also the prevalence of depressive disorders.

The one depression measure that was used concurrently with the PHQ-9 in our subjects was the 5-item mental health scale of the SF-20, also known as the Mental Health Inventory (MHI-5). PHQ-9 scores were strongly correlated with MHI-5 scores in our subjects (Table 4 and Figure 1). Berwick et al. used ROC analysis to determine how well the MHI-5 and several other measures discriminated between patients with and without major depression.[14] In their study, the area under the curve (AUC) was 0.89 for the MHI-5, 0.90 for the longer MHI-18, 0.89 for the 30-item General Health Questionnaire, and 0.80 for the 28-item Somatic Symptom Inventory. In our study, the AUC for major depression was 0.95 for the PHQ-9 and 0.93 for the MHI-5. It is unlikely that other depression-specific measures would be significantly better than the PHQ-9 since an AUC of 1.0 represents a perfect test.

A particularly important characteristic of a severity measure is its sensitivity to change over time. In other words, how precisely do declining or rising scores on the measure reflect improving or worsening depression in response to effective therapy or natural history? Although an exhaustive review of depression measures is beyond the scope of this paper but can be found elsewhere,[4,12] a brief discussion of selected measures is warranted. The Hamilton

Rating Scale for Depression has been the criterion standard outcome measure in clinical trials, but it can require 15 to 30 minutes of clinician time to administer and is therefore not feasible in many practice settings. The HAM-D is also rather complicated to score and requires substantial training in order to get reasonable inter-rater agreement. The Montgomery-Asberg Depression Rating Scale is about half as long as the HAM-D and probably just as sensitive to change.[15,16] Like the HAM-D, however, the Montgomery-Asberg scale must be administered by a clinician with special training and still is moderately time intensive. Several self-administered scales—the 21-item Beck Depression Inventory and the 20-item Zung Self-Rating Depression Scale—also have been used as outcome measures but may be somewhat less sensitive to change than the HAM-D.[17] The SCL-20 has been used as an outcome measure in primary care clinical trials,[18–20] although published evidence on its sensitivity to change as well as other psychometric characteristics is limited. Epidemiological and clinical studies have established the 20-item CES-D as a valid measure for identifying depression, but there is less information regarding its sensitivity to change.

In summary, there appear to be many comparable measures for identifying depression,[1,2,4,12] including a number of self-administered scales. In contrast, it is less clear what the optimal measure for monitoring response to treatment may be, especially outside the setting of a clinical trial. Sensitivity to change is clearly a necessary feature, but other pragmatic considerations include the number of items, time required for completion, mode of administration (self-rating vs interviewer-administered scale), complexity of scoring, inter-rater agreement, and special training requirements. The specific items included in the scale are another factor. One advantage of the PHQ-9 is its exclusive focus on the 9 diagnostic criteria for DSM-IV depressive disorders. On the other hand, some may argue that instruments including symptoms not in the DSM-IV criteria (e.g., loneliness, hopelessness, and anxiety) may have additional value to the clinician. At the same time, it is possible that such scales are less specific for major depression and other mood disorders and may discriminate less accurately depression from anxiety or even general psychological distress.

The major limitation of our study is its cross-sectional design. While our large sample establishes the construct and criterion validity of the PHQ-9, longitudinal studies are needed to establish its sensitivity to change. This will require the completion of several large ongoing clinical trials using the PHQ-9 in parallel with the HAM-D or other established outcome measures. It will also be useful to define the threshold that represents an adequate clinical response. A preliminary approach would be to consider a PHQ-9 score less than 10 and a 50% decline from the pretreatment score as clinically significant improvement. While any proposed threshold requires prospective verification, this approach would be consistent with that established for the HAM-D. Other study limitations are that validation was based on telephone rather than face-to-face interviews and the time for patients to complete the PHQ-9 was not determined.

Detecting depression and initiating treatment are necessary but often insufficient steps to improve outcomes in primary care.[21] Monitoring clinical response to therapy is also critical. Multiple studies have shown that monitoring is often inadequate, resulting in clinician failure to detect medication noncompliance, increase the antidepressant dosage, change or augment pharmacotherapy, or add psychotherapy as needed.[21,22] Having a simple self-administered measure to complete either in the clinic or by telephone administration (e.g., nurse administration[23] or interactive voice recording[24]) would save clinicians the time needed to inquire about the presence and severity of each of the 9 DSM-IV symptoms to assess outcomes.

Brief measures are more likely to be used in the busy setting of clinical practice. For example, many practitioners have found it more feasible to use the 4-item CAGE questionnaire than a number of longer alcohol screening measures. Of note, as few as 1 or 2 questions have demonstrated a high sensitivity in screening for major depression.[2,25] Brevity is just as likely to be a valued attribute when it comes to assessing depression severity as it is when establishing depressive diagnoses. Brevity coupled with its construct and criterion validity makes the PHQ-9 an attractive, dual-purpose instrument for making diagnoses and assessing severity of depressive disorders. If the PHQ-9 proves sensitive to change in clinical trials, it could also be a useful measure for monitoring outcomes of depression therapy.

The development of the PHQ-9 was underwritten by an educational grant from Pfizer US Pharmaceuticals, New York, NY. PRIME-MD is a trademark of Pfizer Copyright held by Pfizer.

REFERENCES

1. Mulrow CD, Williams JW, Gerety MB, Ramirez G, Montiel OM, Kerber C. Case-finding instruments for depression in primary care settings. Ann Intern Med. 1995;122:913–21.
2. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression: two questions are as good as many. J Gen Intern Med. 1997;12:439–45.
3. Keller MB, Kocsis JH, Thase ME, et al. Maintenance phase efficacy of sertraline for chronic depression: a randomized controlled trial. JAMA. 1998;280:1665–72.
4. McDowell I, Kristjansson E, Newell C. Depression. In: McDowell I, Newell C, eds. Measuring Health: A Guide to Rating Scales and Questionnaires. 2nd ed. New York, NY: Oxford University Press; 1996:238–86.
5. Spitzer RL, Kroenke K, Williams JBW. Patient Health Questionnaire Study Group. Validity and utility of a self-report version of PRIME-MD: the PHQ Primary Care Study. JAMA. 1999;282:1737–44.
6. Spitzer RL, Williams JBW, Kroenke K, et al. Validity and utility of the Patient Health Questionnaire in assessment of 3000 obstetric-gynecologic patients: the PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. Am J Obstet Gynecol. 2000;183:759–69.
7. Stewart AL, Hays RD, Ware JE. The MOS Short-Form General Health Survey: reliability and validity in a patient population. Med Care. 1988;26:724–32.

8. Spitzer RL, Williams JBW, Gibbon M, First MB. The structured clinical interview for DSM-III-R (SCID). Arch Gen Psychiatry. 1992;49:624–9.
9. Spitzer RL, Williams JBW, Kroenke K, et al. Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. JAMA. 1994;272:1749–56.
10. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A Basic Science for Clinical Medicine. 2nd ed. Boston, MA: Little, Brown and Company; 1991:1–441.
11. Murphy JM, Berwick DM, Weinstein MC, et al. Performance of screening and diagnostic tests: application of receiver operating characteristic analysis. Arch Gen Psychiatry. 1987;44: 550–5.
12. Pasacreta JV. Measuring depression. In: Frank-Stromborg M, Olsen SJ, eds. Instruments for Clinical Health-Care Research. 2nd Ed. Sudbury, MA: Jones and Bartlett Publishers; 1997: 342–630.
13. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989;27:S178–89.
14. Berwick DM, Murphy JM, Goldman PA, Ware JE, Barsky AJ, Weinstein MC. Performance of a five-item mental health screening test. Med Care. 1991;29:169–76.
15. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979;134:382–9.
16. Davidson J, Turnbull CD, Strickland R, et al. The Montgomery-Asberg Depression Scale: reliability and validity. Acta Psychiatr Scand. 1986;73:544–8.
17. Lambert MJ, Hatch DR, Kingston MD, et al. Zung, Beck, and Hamilton rating scales as measures of treatment outcome: a meta-analytic comparison. J Consult Clin Psychol. 1986;54:54–9.
18. Katon W, Robinson P, Von Korff M, et al. A multifaceted intervention to improve treatment of depression in primary care. Arch Gen Psychiatry. 1996;53:924–32.
19. Katon W, Von Korff M, Lin E, et al. Collaborative management to achieve treatment guidelines: impact on depression in primary care. JAMA. 1995;273:1026–31.
20. Williams JW, Barrett J, Oxman T, et al. Treatment of dysthymia and minor depression in primary care: a randomized controlled trial in older adults. JAMA. 2000;284:1519–26.
21. Kroenke K, Taylor-Vaisey A, Dietrich AJ, Oxman TE. Interventions to improve provider diagnosis and treatment of mental disorders in primary care: a critical review of the literature. Psychosomatics. 2000;41:39–52.
22. Simon GE. Can depression be managed appropriately in primary care? J Clin Psychiatry. 1998;59(suppl 2):3–8.
23. Hunkeler EM, Meresman J, Hargreaves WA, et al. Efficacy of nurse telehealth care and peer support in augmenting treatment of depression in primary care. Arch Fam Med. 2000;9:700–8.
24. Kobak KA, Taylor LH, Dottl SL, et al. A computer-administered telephone interview to identify mental disorders. JAMA. 1997;278:905–10.
25. Williams JW, Mulrow CD, Kroenke K, et al. Case-finding for depression improves patient outcomes: results from a randomized trial in primary care. Am J Med. 1999;106:36–43.

APPENDIX

Nine-symptom Checklist

Name _____ Date _____

Over the *last 2 weeks*, how often have you been bothered by any of the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

(For office coding: Total Score _____ = _____ + _____ + _____)

If you checked off *any* problems, how *difficult* have these problems made it for you to do your work, take care of things at home, or get along with other people?

Not difficult at all	Somewhat difficult	Very difficult	Extremely difficult
☐	☐	☐	☐