MEDICAL CARE
Volume 41, Number 11, pp 1284–1292

The Patient Health Questionnaire-2

Validity of a Two-Item Depression Screener

KURT KROENKE, MD,* ROBERT L. SPITZER, MD,† AND JANET B. W. WILLIAMS, DSW†

BACKGROUND. A number of self-administered questionnaires are available for assessing depression severity, including the 9-item Patient Health Questionnaire depression module (PHQ-9). Because even briefer measures might be desirable for use in busy clinical settings or as part of comprehensive health questionnaires, we evaluated a 2-item version of the PHQ depression module, the PHQ-2.

METHODS. The PHQ-2 inquires about the frequency of depressed mood and anhedonia over the past 2 weeks, scoring each as 0 ("not at all") to 3 ("nearly every day"). The PHQ-2 was completed by 6000 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-2 depression severity increased from 0 to 6, there was a substantial decrease in functional status on all 6 SF-20 subscales. Also, symptom-related difficulty, sick days, and healthcare utilization increased. Using the MHP reinterview as the criterion standard, a PHQ-2 score ≥3 had a sensitivity of 83% and a specificity of 92% for major depression. Likelihood ratio and receiver operator characteristic analysis identified a PHQ-2 score of 3 as the optimal cutpoint for screening purposes. Results were similar in the primary care and obstetrics–gynecology samples.

CONCLUSION. The construct and criterion validity of the PHQ-2 make it an attractive measure for depression screening.

Key words: Depression; diagnosis; screening; mental disorders; functional status (Med Care 2003;41:1284–1292)

Depression is a prevalent and disabling condition in the general medical setting. Although many patients with depression receive care exclusively in the primary care rather than mental health sector, up to half of depression cases in primary care go unrecognized.[1,2] The U.S. Preventive Services Task Force recently concluded that there is sufficient evidence to recommend periodic screening for depression.[1] Numerous well-validated questionnaires are available for depression screening and are similar in terms of their operating characteristics as case-finding instruments.[2–4] One particularly popular instrument is the 9-item depression module of the Patient

From the *Regenstrief Institute for Health Care and Department of Medicine, Indiana University, Indianapolis, Indiana.

From the †New York State Psychiatric Institute and Department of Psychiatry, Columbia University, New York, NY.

The development of the PHQ-2 was underwritten by an educational grant from Pfizer US Pharmaceuticals Inc., New York, NY. PRIME-MD is a trademark of Pfizer Inc. Copyright held by Pfizer Inc.

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

For a complimentary copy of PHQ materials that can be reproduced, e-mail Robert Spitzer at rls8@columbia.edu

Health Questionnaire, the PHQ-9. Validated in 6000 patients, the PHQ-9 serves as both a depression severity measure as well as a diagnostic instrument for the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (DSM-IV), depressive disorders.[5]

However, even shorter measures could be desirable in some circumstances. First, the busy nature and competing demands of primary care practice make efficiency a particularly important attribute of any new measure.[6–8] Second, depression is only one of the many disorders for which screening in primary care is encouraged. Thus, brief depression measures could be incorporated as part of comprehensive health questionnaires administered to either patients new to a practice or established patients on a periodic basis. Keeping to a minimum the number of items asked about a single disorder is an important factor to maintain a reasonable length for such questionnaires. The same might be true for research studies in which depression is a secondary rather than primary variable, and asking a few rather than many items can reduce respondent burden.

Therefore, we examined the operating characteristics of 2 items from the PHQ-9, depressed mood and anhedonia, which we call the PHQ-2. Previous research has shown that a single question about depressed mood has a sensitivity of 85% to 90% for major depression,[3,9] and adding a second question about anhedonia increases the sensitivity to 95%.[3] The PHQ-2 asks respondents to estimate the frequency of these 2 symptoms over the past 2 weeks with 4 response options ranging from "not at all" to "nearly everyday." Data are analyzed from the 2 major PHQ studies involving 3000 patients in 8 primary care clinics and 3000 patients in 7 obstetrics–gynecology clinics.[10,11] Our aims were to assess the criterion and construct validity of the PHQ-2.

Methods

Description of the PHQ-9 and PHQ-2

The PHQ-9 is the 9-item depression module from the full PHQ.[5] Major depressive disorder 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 one of the symptoms is depressed mood or anhedonia. "Other depressive disorder" 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 one 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 as one of the diagnostic criteria for depressive disorders if present at all, regardless of duration. Like with the original PRIME-MD (Pfizer Inc., New York, NY), the clinician is expected to rule out physical causes of depression, normal bereavement, and history of a manic episode.

The PHQ-2 includes the first 2 items of the PHQ-9. The stem question is, "Over the last 2 weeks, how often have you been bothered by any of the following problems?" The 2 items are "little interest or pleasure in doing things" and "feeling down, depressed, or hopeless." For each item, the response options are "not at all," "several days," "more than half the days," and "nearly everyday," scored as 0, 1, 2, and 3, respectively. Thus, the PHQ-2 score can range from 0 to 6.

PHQ Study Samples and Procedures

From May 1997 to November 1998, 3890 patients, 18 years or older, were invited to participate in the PHQ Primary Care Study.[10] There were 190 who declined to participate, 266 who started but did not complete the questionnaire (often because there was an inadequate time before seeing their physician), and 434 whose questionnaires were not entered into the dataset because the equivalent of approximately one page (20 items) was not completed. This resulted in the 3000 primary care patients reported here (1422 from 5 general internal medicine clinics and 1578 from 3 family practice clinics). From May 1997 to March 1999, 3636 patients, 18 years or older, were approached to participate in the PHQ Obstetrics–Gynecology Study.[11] 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 dataset because the equivalent of approximately one page was not completed. This resulted in the 3000 subjects from 7 obstetrics–gynecology sites. The 2 PHQ studies enrolled patients from 8 states (New York, Massachusetts, Utah, Pennsylvania, Texas, Virginia, Ohio, and Wisconsin) plus the District of Columbia.

All sites used 1 of 2 subject selection methods to minimize sampling bias: either consecutive pa-

tients for a given clinic session or every nth patient until the intended quota for that session was achieved. Patient characteristics are summarized in detail elsewhere.[5] Briefly, the primary care sample was 66% women, 21% minority in terms of race/ethnicity, and had a mean age of 46 years. The obstetrics–gynecology sample was 100% women, 61% minority, and had a mean age of 31 years. The Institutional Review Board at each site approved the study protocol.

Before seeing the physician, all patients completed the full self-administered PHQ. Also, they completed the Medical Outcomes Study Short-Form General Health Survey (SF-20).[12] The SF-20 measures functional status in 6 domains (all scores from 0–100, 100 = best health). Additionally, patients estimated the number of physician visits and disability days during the past 3 months.

Mental Health Professional Validation Interviews

To determine the agreement of PHQ diagnoses with those of mental health professionals (MHPs), midway through the PHQ Primary Care Study, an 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 one 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 (structured clinical interview for DSM-III-R)[13] and diagnostic questions from the PRIME-MD, are described in the original PRIME-MD report.[14] 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.

Analysis

For analyses assessing the operating characteristics of various PHQ-2 cutpoints, 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-2 score. Besides calculating sensitivity and specificity of the PHQ-2 over various intervals, we also determined likelihood ratios[15] and conducted receiver operating characteristic (ROC) curve analysis[16] as quantitative methods for combining sensitivity and specificity into a single metric.

Construct validity of the PHQ-2 as a measure of depression severity was assessed by examining functional status (the 6 SF-20 scales), disability days, symptom-related difficulty, and healthcare utilization (clinic visits) over the various PHQ-2 scores. Analysis of covariance was used with PHQ-2 score as the independent variable and adjusting for age, sex, race, education, study site, and number of physical disorders. Bonferroni's correction was used to adjust for multiple comparisons.

Decrements in SF-20 scores were also examined in terms of effect size, which is the difference in mean SF-20 scores, expressed as the number of standard deviations, between each PHQ-2 score and the reference group. The reference group was subjects with a PHQ-2 score of 0, and the standard deviation used was that of the entire sample.

Results

Distribution of PHQ-2 Scores According to Depression Diagnostic Status

Table 1 shows the distribution of PHQ-2 scores according to depression diagnostic status in the 580 patients interviewed by an MHP who was blinded to the PHQ-2 results. Of the 41 subjects who had major depressive disorder, 93% endorsed at least some depressed mood (1 or greater) and 95% endorsed at least some anhedonia. The majority of patients (95%) with no depressive disorder had a PHQ-2 score less than 3, whereas most patients (83%) with major depression had scores of 3 or greater. Patients with other (ie, nonmajor) depressive disorder exhibited more heterogeneity in their PHQ-2 scores.

Criterion Validity of the PHQ-2 Assessed by Mental Health Professional Interview

Table 2 displays the sensitivity, specificity, positive predictive value, and likelihood ratios for

TABLE 1. Frequency Distribution of Depressed Mood and Anhedonia Items and PHQ-2 Score in a Subset of 580 Patients Who Had an MHP Interview in PHQ Primary Care Study*

	Major Depression (N = 41)		Other Depression (N = 65)		No Depression (N = 474)	
Item or PHQ-2 Score	N	%	N	%	N	%
Depressed mood						
0 (not at all)	3	7.3	14	21.5	338	71.3
1 (several days)	7	17.1	23	35.4	114	24.1
2 (more than half)	16	39.0	17	26.2	13	2.7
3 (nearly every day)	15	36.6	11	16.9	9	1.9
Anhedonia						
0 (not at all)	2	4.9	16	24.6	364	76.8
1 (several days)	8	19.5	25	38.5	89	18.8
2 (more than half)	10	24.4	13	20.0	16	3.4
3 (nearly every day)	21	51.2	11	16.9	5	1.1
PHQ-2 Score						
0	1	2.4	9	13.8	310	65.4
1	2	4.9	7	10.8	71	15.0
2	4	9.8	17	26.2	71	15.0
3	4	9.8	8	12.3	10	2.1
4	8	19.5	13	20.0	6	1.3
5	11	26.8	9	13.8	5	1.1
6	11	26.8	2	3.1	1	0.2

*Depression diagnostic status was determined in 580 primary care patients by having a mental health professional (MHP) who was blinded to the PHQ-2 score administer a structured psychiatric interview.

PHQ, Patient Health Questionnaire.

different PHQ-2 scores in diagnosing depressive disorders in the 580 patients who had an MHP interview. In this sample with a 7% prevalence of major depressive disorder, the positive predictive value for major depressive disorder ranged from 21% for a PHQ-2 cutpoint of 2 to 56% for a

TABLE 2. Operating Characteristics of PHQ-2 as a Screener for Depressive Disorders in 580 Patients Who Had an Independent Mental Health Professional Interview†

	Major Depressive Disorder (N = 41)				Any Depressive Disorder (N = 106)			
PHQ-2	Sensitivity	Specificity	Positive Predictive Value	Likelihood Ratio	Sensitivity	Specificity	Positive Predictive Value	Likelihood Ratio
1	97.6	59.2	15.4	0.3	90.6	65.4	36.9	0.6
2	92.7	73.7	21.1	0.6	82.1	80.4	48.3	1.3
3	82.9	90.0	38.4	2.9	62.3	95.4	75.0	5.4
4	73.2	93.3	45.5	5.5	50.9	97.9	81.2	15.7
5	53.7	96.8	56.4	10.3	31.1	98.7	84.6	17.9
6	26.8	99.4	78.6	48.2	12.3	99.8	92.9	58.1

†Sensitivity, specificity, and positive predictive value refers to a *threshold* PHQ-2 score (ie, all subjects with that score or higher), whereas likelihood ratio refers to a *discrete* PHQ-2 score (ie, only subjects with that specific score). For example, 82.9% of patients with major depressive disorder have a PHQ-2 score of 3 or greater (sensitivity), 90% of patients without major depressive disorder have a score of less than 3 (specificity), 38.4% of patients with a score of 3 or greater have major depressive disorder (positive predictive value), and a score of 3 is 2.9 times more likely in patients with than without major depressive disorder (likelihood ratio).

PHQ, Patient Health Questionnaire.

cutpoint of 5. The positive predictive value for any depressive disorder (which had a prevalence of 18%) ranged from 48% for a PHQ-2 cutpoint of 2 to 85% for a cutpoint of 5. At a cutpoint of 3 or higher, the PHQ-2 had a likelihood ratio for major depression of 2.92, nearly identical to the overall likelihood ratio of 2.86 reported for 9 other depression case finding instruments in a meta-analysis of the literature.[2] Regarding concordance with the MHP interview, a PHQ-2 cutpoint of 3 or greater was comparable to the PHQ-9 diagnostic algorithm for any depressive disorder (kappa of 0.62 vs. 0.58) as well as major depressive disorder (kappa of 0.48 vs. 0.54).

ROC analysis showed that the area under the curve (AUC) for the PHQ-2 in diagnosing major depressive disorder was 0.93 (vs. 0.95 for the longer PHQ-9). The AUC of the PHQ-2 for diagnosing any depressive disorder was 0.90 (vs. 0.92 for the PHQ-9). Although the AUC for both major and any depressive disorder was similar for women and men, age had a modest effect. The AUC for major depressive disorder was somewhat greater in subjects less than 60 years compared with those 60 years and older (0.94 vs. 0.86), whereas for any depressive disorder, younger subjects had a lower AUC (0.88 vs. 0.95).

Construct Validity of the PHQ-2 Assessed by Functional Status and Other Measures

As shown in Table 3, there was a strong association between increasing PHQ-2 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-2 scores were greatest for the scales that previous studies have shown should be most strongly related to depression, ie, mental health, followed by social, overall, and role functioning, with a lesser relationship to pain and physical functioning.[14] Third, most pairwise comparisons within each SF-20 scale between successive PHQ-2 levels were highly significant.

Figure 1 illustrates graphically the relationship between increasing PHQ-2 scores and worsening functional status. Decrements in SF-20 scores are shown in terms of effect size. Effect sizes of 0.5 and 0.8 are typically considered moderate and large between-group differences, respectively.[17] Figure 1 shows effect sizes for the primary care sample; results for the obstetrics–gynecology sample (not displayed) were similar.

When the PHQ-2 was examined as a continuous variable, its strength of association with the SF-20 scales was concordant with the pattern seen in Figure 1. In both the primary care and obstetrics–gynecology samples, the PHQ-2 correlated most strongly with mental health (0.70 and 0.63), followed by general health perceptions (0.47 and 0.46), social functioning (0.46 and 0.36), physical functioning (0.37 and 0.36), role functioning (0.37 and 0.29), and bodily pain (0.26 and 0.31).

Table 4 shows the association between PHQ-2 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, healthcare utilization, and symptom-related difficulty in activities and relationships.

Discussion

Our study provides strong evidence for the validity of the PHQ-2 as a brief depression screening measure. Criterion validity was demonstrated by the fact that the operating characteristics of the PHQ-2 compared favorably with an independent MHP interview in a sample of 585 patients. Construct validity was established by the strong association between PHQ-2 scores and functional status, disability days, and symptom-related difficulty. The sample of 6000 patients from 15 geographically dispersed clinics as well as the similarity of findings in 2 different patient populations (primary care and obstetrics–gynecology) enhances the generalizability of our findings.

The PHQ-9 would still be the preferred instrument when the intent is either to definitively diagnose depressive disorders or to assess depression outcomes in response to treatment. This is because the PHQ-9 includes all 9 symptom criteria necessary for establishing DSM-IV depressive disorder diagnoses and provides a wider range of depressive symptom severity scores (0–27) compared with the PHQ-2 (0–6). However, in many settings, the purpose is not to establish final diagnoses or to monitor depression severity, but rather to screen for depression in a "first step" approach. Even briefer versions might be desirable

TABLE 3. Relationship between PHQ-2 Depression Score and SF-20 Health-Related Quality-of-Life Scales in Patients in the PHQ Primary Care (n = 3000) and PHQ Obstetrics–Gynecology (n = 3000) Studies†

	Mental		Social		Role		General		Pain		Physical	
PHQ-2 Score	Primary Care	Ob-Gyn	Primary Care	Ob-Gyn	Primary Care	Ob-Gyn	Primary Care	Ob-Gyn	Primary Care	Ob-Gyn	Primary Care	Ob-Gyn
	Mean SF-20 scale score											
0	82.1	82.2	90.9	91.7	84.1	88.3	68.8	74.9	65.1	73.7	81.5	85.3
1	69.8	73.8	85.4	87.1	76.9	84.4	58.8	65.7	59.3	65.8	75.1	81.1
2	58.7	62.4	74.8	80.8	60.3	76.9	48.6	57.0	52.9	58.8	68.5	77.3
3	52.6	53.6	60.6	74.1	50.2	70.7	38.5	50.5	46.9	54.7	61.8	76.7
4	46.5	50.2	57.6	67.1	51.3	57.7	39.5	43.4	49.7	53.5	61.4	68.0
5	39.2	43.9	59.7	62.0	39.5	53.1	35.7	41.8	46.2	43.8	58.5	63.5
6	33.3	40.3	47.4	54.7	37.0	49.3	30.2	36.2	43.7	48.0	57.8	65.2

†SF-20 scores are adjusted for age, sex, race, education, study site, and number of physical disorders.

To simplify this table, standard deviations and *P* values are not shown. However, most pair-wise 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.

PHQ, Patient Health Questionnaire; SF-20, 20-item Short- Form General Health Survey.

when the aim is to include just a few depression questions in multipurpose health questionnaires. The U.S. Preventive Services Task Force recently recommended depression screening as part of routine care.[1] However, brevity is essential to accomplish this in the busy general medical setting where patient volume is high, most visits are brief, and depression is simply one of many conditions that the primary care clinician is responsible for recognizing and managing.[6–8]

Others have shown that 1 or 2 questions about depressed mood and, possibly, anhedonia are quite sensitive as a first-stage depression screening procedure.[2–4,9,18] Whereas Whooley and colleagues also found that the 2 depression items of the original PRIME-MD performed similarly to

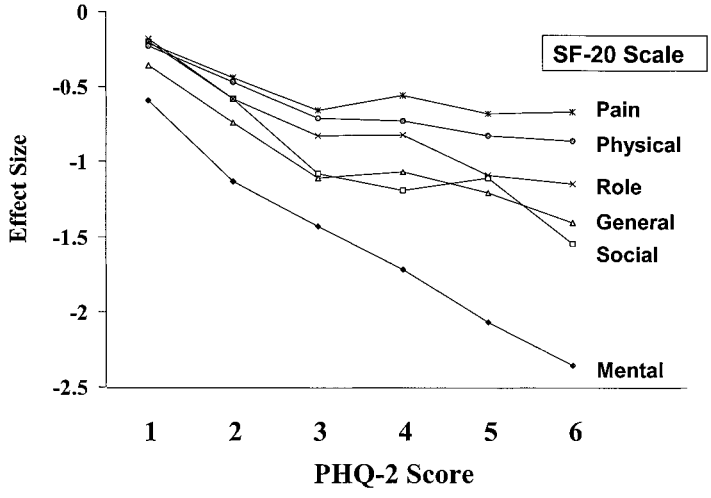

FIG. 1. Relationship between depression severity as measured by the PHQ-2 and decline in functional status as measured by the 6 subscales of the SF-20. The decrement in SF-20 scores is shown as the difference between each PHQ-2 score and the reference group (ie, those with a PHQ-2 of 0). Effect size is the difference in group means divided by the standard deviation of the entire sample.

TABLE 4. Relationship between PHQ-2 Depression Score and Disability Days, Symptom-Related Difficulty, and Clinic Visits in the PHQ Primary Care, and Obstetrics–Gynecology Studies

	Mean Disability Days*		Mean Physician Visits*		Symptom-Related Difficulty (%)†	
PHQ-2 Score	Primary Care	Obstetrics-Gynecology	Primary Care	Obstetrics-Gynecology	Primary Care	Obstetrics-Gynecology
0	2.7	2.3	1.1	0.9	2.1%	1.4%
1	4.6	3.0	1.3	1.0	6.4%	2.2%
2	9.0	6.0	1.9	1.2	11.9%	3.7%
3	9.4	6.8	1.9	2.0	24.5%	17.1%
4	12.8	12.8	2.5	1.4	31.3%	25.0%
5	16.7	13.3	2.3	1.8	53.1%	36.6%
6	25.0	16.8	3.0	1.9	57.0%	48.4%

*_Disability days_ refers to number of days in past 3 months that patients' symptoms interfered with their usual activities. _Physician visits_ refers to past 3 months also. Both are self-reported. Means are also adjusted for age, sex, 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." _Symptom-related difficulty_ in this table refers to those patients reporting "very" or "extremely" difficult.

To simplify this table, standard deviations and P values are not shown. However, 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.

PHQ, Patient Health Questionnaire.

longer case-finding instruments,[3] our findings build on this earlier work for 4 important reasons. First, our PHQ studies included 6000 patients from multiple clinics representing a more diverse population, whereas Whooley et al. studied 536 patients who were mostly male (97%) veterans drawn from a single clinic. Second, the original PRIME-MD evaluated by Whooley et al. was designed as a 2-stage procedure in which positive responses on the patient questionnaire prompted a structured interview using a clinician evaluation guide. To maximize sensitivity, the timeframe for depressed mood and anhedonia was the past month and the response options were simply yes or no. Because the PHQ was intended to be exclusively a patient self-report version of the PRIME-MD, 2 important modifications were made: the timeframe focused on the past 2 weeks and the response options were expanded to 4 to better delineate the number of days patients were bothered by depressed mood and/or anhedonia. This makes the one-stage PHQ more consistent with DSM-IV criteria for depressive disorders and therefore greatly improves the specificity of the PHQ-2 compared with the 2 items of the PRIME-MD with only a modest decline in sensitivity. Specificity is an important consideration if depression screening becomes more widespread, because a large number of false-positives would be difficult to handle efficiently in the context of the large patient volume, short visits, and competing demands of primary care.[6–8] Third, the characteristic of better discriminating between depressed and nondepressed patients is exemplified by the PHQ-2's higher AUC for major depression of 0.93 compared with 0.82 for the 2-item PRIME-MD as reported by Whooley et al. Because this comparison is drawn from 2 different studies (albeit both primary care), a head-to-head comparison of the operating characteristics of the PHQ-2 versus the 2-item PRIME-MD would optimally be performed in the same patient population. Fourth, the PHQ is rapidly replacing the original PRIME-MD in both research and clinical settings, so understanding the operating characteristics of the PHQ-2 as a depression screener is important for pragmatic reasons.

The operating characteristics of the PHQ-2 displayed at various cutpoints in Table 2 compare favorably to 9 other case-finding instruments for depression in primary care that have an overall sensitivity of 84%, a specificity of 72%, and a positive likelihood ratio of 2.86.[2] At a cutpoint of 3, the PHQ-2 has a sensitivity of 83%, a specificity of 90%, and a positive likelihood ratio of 2.9. Likewise, the positive predictive value of the PHQ-2

for major depression in our sample in which the prevalence of major depression was 7% (similar to other outpatient samples) ranged from 20% to 45% as the cutpoint was varied from 2 to 5. This predictive value is similar to other instruments. Of note, predictive value is related not only to a measure's sensitivity and specificity, but also to the prevalence of depressive disorders.

The one depression measure that was used concurrently with the PHQ-2 in our subjects was the 5-item mental health scale of the SF-20, also known as the Mental Health Inventory (MHI-5). PHQ-2 scores were strongly correlated with MHI-5 scores in both the primary care (r = .70) and obstetrics–gynecology (r = .63) samples, an association clearly illustrated in Table 3 and Figure 1. Berwick and colleagues used ROC analysis to determine how well the MHI-5 and several other measures discriminated between patients with and without major depression.[19] In their study, the 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 both the PHQ-2 and the MHI-5. Because an AUC of 1.0 signifies a perfect test, it is unlikely that other depression measures are diagnostically superior.

Several caveats should be mentioned. First, screening with the PHQ-2 is only a first step. Patients who screen positive should be further evaluated with the PHQ-9, other diagnostic instruments, or direct interview to determine whether they meet criteria for a depressive disorder. High scores on the PHQ-2 alone would typically not be a sufficient basis to initiate treatment without diagnostic confirmation. Second, subjects in our study completed the PHQ-2 as part of the PHQ-9. An alternative design would be to administer the PHQ-2 to one group of subjects and the PHQ-9 to a second comparable group; establishing similar operating characteristics with this method would further validate the PHQ-2 as a stand-alone depression screener. Third, picking the optimal PHQ-2 cutpoint, like with any measure, is a trade-off between sensitivity and specificity. Although our analysis suggests a cutpoint of 3 provides a reasonable compromise, a cutpoint of 2 would enhance sensitivity, whereas a cutpoint of 4 would improve specificity. One must be cautious about overemphasizing sensitivity with depression screening in primary care, partly because of the high volume of patients in primary care and partly because a very sensitive cutpoint coupled with a 5% to 10% prevalence of major depression means that most patients screening positive are false-positive cases.[2] In settings in which the prevalence of depression is much higher (eg, psychiatric settings or hospitalized patients), a different cutpoint might be considered.

A third caveat is that depression screening by itself is not enough.[20,21] Adequate follow up, depression severity monitoring, and access when needed to MHPs are required to detect medication noncompliance, increase the antidepressant dosage, change or augment pharmacotherapy, or add psychotherapy as needed.[22–24] Even the U.S. Preventive Services Task Force concludes that depression screening is only effective if coupled with systems changes to appropriately diagnose and treat depression.[1] In this sense, depression is no different than many other medical disorders such as diabetes, hypertension, and asthma, in which detection must be combined with initial patient education and activation and ongoing disease monitoring and management, with a partnership forged between the patient, the primary care provider, and the specialist.

Asking about depressed mood and anhedonia would typically not require a paper questionnaire, unless the items are included as part of a longer health survey. Most clinicians could easily remember these 2 core symptoms. The PHQ-2 response options simply allow patients to grade the amount of time they have been bothered by either symptom in the past 2 weeks, a core feature of the DSM-IV criteria and one that distinguishes fleeting from more persistent mood symptoms. Being bothered by either depressed mood or anhedonia "nearly everyday" or one symptom "more than half the days" and the other symptom "several days" would result in a PHQ-2 score of 3. This score (or a lower score if other "red flags" of depression are present[8]) could trigger administration of the full PHQ-9.

Brevity is a particularly attractive feature of measures intended for use in clinical practice. In the words of Mies van de Rohe, the modern architect, "less is more." For example, one reason for the popularity of the 4-item CAGE questionnaire (a questionnaire for alcoholism evaluation)[25] could be its brevity compared with longer alcohol screening measures. Likewise, simplifying depression screening to 2 questions enhances routine

inquiry about the most prevalent and treatable mental disorder in primary care.

References

1. **U.S. Preventive Services Task Force**. Screening for depression: recommendations and rationale. Ann Intern Med 2002;136:760–764.

2. **Mulrow CD, Williams JW, Gerety MB, et al.** Case-finding instruments for depression in primary care settings. Ann Intern Med 1995;122:913–921.

3. **Whooley MA, Avins AL, Miranda J, et al.** Case-finding instruments for depression: two questions are as good as many. J Gen Intern Med 1997;12:439–445.

4. **Williams JWJ, Noel PH, Cordes JA, et al.** Is this patient clinically depressed? JAMA 2002;287:1160–1170.

5. **Kroenke K, Spitzer RL, Williams JBW.** The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16:606–613.

6. **Klinkman MS.** Competing demands in psychosocial care: a model for the identification and treatment of depressive disorders in primary care. Gen Hosp Psychiatry 1997;19:98–111.

7. **Williams JW Jr.** Competing demands: does care for depression fit in primary care? J Gen Intern Med 1998;13:137–139.

8. **Kroenke K.** Discovering depression in medical patients: reasonable expectations. Ann Intern Med 1997;126:463–465.

9. **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.

10. **Spitzer RL, Kroenke K, Williams JBW, et al.** Validity and utility of a self-report version of PRIME-MD: the PHQ primary care study. JAMA 1999;282:1737–1744.

11. **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;183:759–769.

12. **Stewart AL, Hays RD, Ware JE.** The MOS Short-Form General Health Survey: reliability and validity in a patient population. Med Care 1988;26:724–732.

13. **Spitzer RL, Williams JBW, Gibbon M, et al.** The Structured Clinical Interview for DSM-III-R (SCID). Arch Gen Psychiatry 1992;49:624–629.

14. **Spitzer RL, Williams JB, Kroenke K, et al.** Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. JAMA 1994;272:1749–1756.

15. **Sackett DL, Haynes RB, Guyatt GH, et al.** Clinical Epidemiology: A Basic Science for Clinical Medicine, 2nd ed. Boston: Little, Brown and Co; 1991.

16. **Murphy JM, Berwick DM, Weinstein MC, et al.** Performance of screening and diagnostic tests: application of receiver operating characteristic analysis. Arch Gen Psychiatry 1987;44:550–555.

17. **Kazis LE, Anderson JJ, Meenan RF.** Effect sizes for interpreting changes in health status. Med Care 1989;27:S178–S189.

18. **Williams JW Jr, Pignone M, Ramirez G, et al.** Identifying depression in primary care: a literature synthesis of case-finding instruments. Gen Hosp Psychiatry 2002;24:225–237.

19. **Berwick DM, Murphy JM, Goldman PA, et al.** Performance of a five-item mental health screening test. Med Care 1991;29:169–176.

20. **Kroenke K.** Depression screening is not enough. Ann Intern Med 2001;134:418–420.

21. **Valenstein M, Vijan S, Zeber JE, et al.** The cost-utility of screening for depression in primary care. Ann Intern Med 2001;134:345–360.

22. **Kroenke K, Taylor-Vaisey A, Dietrich AJ, et al.** Interventions to improve provider diagnosis and treatment of mental disorders in primary care: a critical review of the literature. Psychosomatics 2000;41:39–52.

23. **Simon GE.** Can depression be managed appropriately in primary care? J Clin Psychiatry 1998;59(suppl 2):3–8.

24. **Oxman TE, Dietrich AJ, Williams JW, et al.** A three component model for reengineering systems for the treatment of depression in primary care. Psychosomatics 2002;43:441–450.

25. **Bush B, Shaw S, Cleary P, et al.** Screening for alcohol abuse using the CAGE questionnaire. Am J Med 1987;82:231–235.