Screening for Depression and Suicidality in a VA Primary Care Setting: 2 Items Are Better Than 1 Item

Kathryn Corson, PhD; Martha S. Gerrity, MD, MPH, PhD; and Steven K. Dobscha, MD

Objective: To evaluate the psychometric properties of a single-item depression screen against validated scoring algorithms for the Patient Health Questionnaire (PHQ) and the utility of these algorithms in screening for depression and suicidality in a Department of Veterans Affairs (VA) primary care setting.

Study Design: Recruitment phase of a randomized trial.

Methods: A total of 1211 Portland VA patients with upcoming primary care clinic appointments were administered by telephone a single item assessing depressed mood over the past year and the PHQ. The PHQ-9 (9 items) encompasses DSM-IV criteria for major depression, the PHQ-8 (8 items) excludes the thoughts of death or suicide item, and the PHQ-2 (2 items) assesses depressed mood and anhedonia. Patients whose responses suggested potential suicidality were administered 2 additional items assessing suicidal ideation. Patients receiving mental health specialty care were excluded.

Results: Using the PHQ-9 algorithm for major depression as the reference standard, the VA single-item screen was specific (88%) but less sensitive (78%). A PHQ-2 score of ≥3 demonstrated similar specificity (91%) with high sensitivity (97%). For case finding, the PHQ-8 was similar to the PHQ-9. Approximately 20% of patients screened positive for moderate depression, 7% reported thoughts of death or suicide, 2% reported thoughts of harming themselves, and 1% had specific plans.

Conclusions: The PHQ-2 offers brevity and better psychometric properties for depression screening than the single-item screen. The PHQ-9 item assessing thoughts of death or suicide does not improve depression case finding; however, one third of patients endorsing this item reported recent active suicidal ideation.

(Am J Manag Care. 2004;10(part 2):839-845)

Depression is common among patients in primary care settings, yet it is underrecognized and undertreated by primary care providers.[1-3] Given the high prevalence, morbidity, and mortality associated with untreated depression, many medical institutions have initiated systematic guideline-based screening programs.[4-6] Widely used screening instruments include the Beck Depression Inventory, the Center for Epidemiologic Studies Depression Screen (CES-D), and the Zung Self-Assessment Depression Scale.[7] Compared with a standardized diagnostic instrument, these screens demonstrate very good sensitivity and fair to good specificity.[8]

Still, administering and evaluating the 20 or more items typically found in measures of depression can be relatively time-consuming, and therefore difficult to integrate into busy primary care practices.[9,10] Thus, shorter instruments have been introduced and tested.[8,11-14] Of note, the recently developed 9-item Patient Health Questionnaire (PHQ-9)[15-17] is increasingly being administered and tested in clinical and research settings.[18-23] The PHQ-9 has good sensitivity (88%) and specificity (88%) for major depression compared with a diagnostic interview conducted by a mental health professional using SCID (Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders, Revised 3rd Edition* [DSM-III-R]) criteria.[17] The PHQ-9 offers concurrent validity with measures of functional impairment, high internal consistency and test-retest reliability, simplicity, and face validity[15-19]; in addition, severity scores may be used to track change over time.[7,16,23-24]

Looking at the shortest possible measures, Whooley et al[13] found that 2 items (measuring depressed mood and anhedonia over the past month) demonstrated excellent sensitivity (96%) but only fair specificity (57%) compared with the Diagnostic Interview Schedule. Kroenke et al[25] tested the validity of the first 2 items (depressed mood and anhedonia over the past 2 weeks) of the PHQ (PHQ-2) in a population of community primary care and obstetrics-gynecology patients. They found that a PHQ-2 score of 3 or higher (PHQ-2 ≥3) had a sensitivity of 83% and specificity of 92% compared with a diagnostic interview by a mental health profes-

From Research Service (KC), Behavior Health and Clinical Neurosciences Division (SKD), and the Division of Hospital and Specialty Medicine (MSG), Portland VA Medical Center, Portland, Ore; and the Department of Psychiatry (KC, SKD) and the Department of Medicine (MSG), Oregon Health & Science University, Portland.

This study was supported by the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service project MHI 20-020-1. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs.

Address correspondence to: Kathryn Corson, PhD, Portland VA Medical Center, PO Box 1034 (P3 DEP-PC), Portland, OR 97207. E-mail: kathryn.corson@med.va.gov.

sional. Also using a diagnostic interview as the criterion, Williams et al[26] reported that the sensitivity and specificity for a single question ("Have you felt depressed or sad much of the time in the past year?") approached that of the CES-D (85% vs 88% and 66% vs 75%, respectively). Although the data of Williams et al suggest that 1 item performs well, the characteristics of their sample—predominantly female and Hispanic—limit generalization to other settings.

In 1999, the Portland Veterans Affairs Medical Center (VAMC) primary care clinics introduced a similar single item ("Have you been depressed or sad most of the past year?") as a routine annual depression screen. In contrast to Williams et al's population, the VA patient population is predominantly male, Caucasian, and older adults.[27] The primary objective of this study was to evaluate the sensitivity and specificity of the single-item screen with the PHQ-9 as the reference standard in a VA primary care clinic. We also sought to estimate the proportion of primary care patients not currently receiving mental health specialty care who would screen positive for depression and possible suicidality.

PATIENTS AND METHODS

Setting

The study was conducted in the Portland VAMC primary care clinics, which include 2 hospital-based and 2 community-based clinics. In 2002, about 23 000 patients were followed in these clinics. Our local population is primarily older (mean age 62 years), Caucasian (87% of patients with recorded ethnicity) men (94%), reflecting national VA demographics. The modal panel size for physicians is 1100-1200 patients; for nurse practitioners and physician assistants, 760-960 patients.

Study Sample and Procedure

In July 2002 we initiated recruitment for a randomized, controlled trial of a low-intensity collaborative intervention for depression in primary care (DEP-PC). All patients screened for participation in DEP-PC between July 2002 and February 2003 were eligible for the current study. Potential participants in DEP-PC were identified by using computerized lists of patients due to see their primary care providers within a month and whose primary care providers (n = 41) were participating in DEP-PC. We excluded patients who had received treatment from a mental health care clinician within the prior 6-month period or who had Alzheimer's disease, cognitive problems, psychotic symptoms, or terminal illness documented in their medical records (the **Figure**).

Patients who met inclusion criteria for DEP-PC were sent a brief letter outlining the study. One to two weeks later, a research assistant telephoned, explained the purpose of the study, and asked permission to continue with a 5-minute telephone interview. Up to 3 call attempts were made to reach each patient. It has been established that depression data collected by telephone are comparable to depression data obtained by in-person interviews.[28] Research assistants were trained in procedures for obtaining clinical assistance for severely depressed or potentially suicidal patients. Investigators contacted patients who expressed active suicidal ideation for assessment and to offer care. The local institutional review board approved the study.

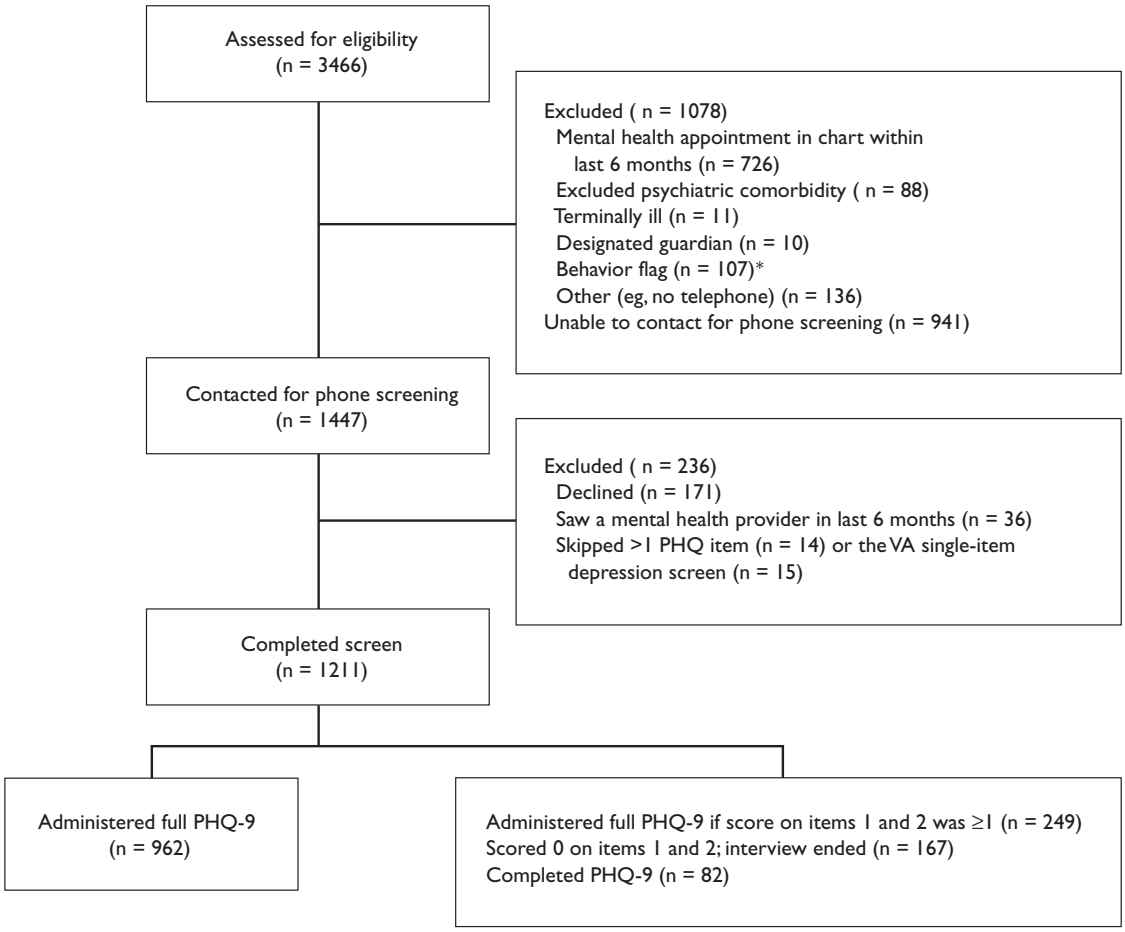
All eligible patients who agreed to be screened for DEP-PC were administered the PHQ and the single-item screen currently used in the primary care clinics. Over the first 5 months of recruitment, 977 patients were screened. Of the 587 patients who answered "not at all" to the first 2 PHQ-9 items (anhedonia and depressed mood), more than half (54%) also answered "not at all" to each of the remaining 7 PHQ items. Moreover, only 3 of 587 (0.5%) patients had PHQ scores suggesting moderate depression (PHQ-9 $\geq$ 10). Therefore, to limit the length of screening calls, we began administering the full PHQ-9 only when patients endorsed at least 1 of the first 2 PHQ items. Those interviewed using this "abbreviated screen" who did not endorse either of the first 2 items (n = 167) received a score of zero and the interview ended.

Over the 7-month study period, 1447 veterans enrolled in the primary care clinics were contacted by phone for screening (the Figure). Of these, 1240 (85.7%) patients completed the screening, 171 (11.8%) declined to be screened, and 36 (2.5%) indicated that they had seen a mental health clinician in the past 6 months. Among the 1240 screened, 14 (0.1%) patients skipped 2 or more PHQ items and 15 (0.1%) patients did not answer the single-item screen, leaving a final sample size of 1211. Veterans screened for DEP-PC were slightly more likely to be Caucasian (93% of patients with recorded ethnicity) and older (mean age 66 years) than veterans in the general primary care population.

Measures

Patient Health Questionnaire-9. The PHQ-9 depression scale is derived from the PRIME-MD,[15-17] a measure of mood, anxiety, alcohol, somatoform, and eating disorders with demonstrated diagnostic and concurrent validity. Patients use an ordinal scale (0 = not at all, 1 = several days, 2 = more than half the days, 3 = nearly every day) to rate the frequency of symptoms of depression over the past 2 weeks. The 9 items are based on the

Figure. Flowchart of Participants



*A "behavior flag" in the medical chart indicates that patient has a specialized care plan due to a history of disruptive behaviors (including drug seeking). PHQ indicates Patient Health Questionnaire.

9 DSM-IV criteria for the diagnosis of depression,[29] and total scores range from 0 to 27. Options for administering and scoring the PHQ include using all 9 items, using the first 8 items (PHQ-8; excludes thoughts of death or suicide item), and using only the first 2 items (PHQ-2; anhedonia and depressed mood items). For classification, either the cut-point system (score of 5-9 = mild, 10-14 = moderate, 15-19 = moderately severe, and 20-27 = severe depression) or the algorithm developed and validated by Spitzer and his colleagues[17] to be congruent with the DSM-IV criteria ("major depression algorithm") can be used. The PRIME-MD also contains an item to assess global functional impairment that can be administered in conjunction with the PHQ as a 10th item. Our screening protocol used the 9-item version, previously validated against clinician interview and measures of functional impairment.[15,17,19]

The last item of the PHQ-9 evaluates the frequency of "thoughts that you would be better off dead or of hurting yourself in some way." We developed 2 additional follow-up questions for patients endorsing this item. The first was designed to clarify whether the patient is experiencing active suicidal ideation ("Are these thoughts that you would be better off dead, or thoughts of hurting or killing yourself?"). The second asks about active planning ("Over the past 2 weeks have you thought about specific ways you might hurt or kill yourself?").

Single-Item Screen. In 1997, the Veterans Health Administration (VHA) released clinical practice guidelines for major depressive disorder, which included annual screening for all general medicine patients.[4] At the Portland VAMC, primary care patients are screened annually for depression unless they are currently

undergoing specialty mental health treatment. The screening item "Have you been depressed or sad most of the past year?" uses a yes/no response format and is based on the single-item tested by Williams and his colleagues.[26]

Statistical Analysis

When a patient skipped a single PHQ item (17/1211, or <1.5%), the omitted value was imputed using mean substitution.[30] Imputed data were not used in the analysis of detection of suicidal ideation. Internal consistency (Cronbach's alpha) was calculated by using data from patients interviewed during the first 5 months of recruitment who answered all 9 items (n = 962). There were no differences between the 962 (79%) patients assessed with the full PHQ and the 249 patients assessed with the abbreviated screen in terms of demographics or depression severity (ie, the proportion in each cohort classified as not depressed, mildly depressed, moderately depressed, etc).

Receiver operating characteristic (ROC) curve analyses comparing patients screened before and after the change in PHQ administration procedure showed no significant differences for the single item, PHQ-2 ≥ 2, or PHQ-2 ≥ 3 when the major depression diagnosis algorithm, PHQ-9 ≥ 10, or PHQ-9 ≥ 15 was used as the reference standard. Thus, the data were combined for all subsequent analyses except PHQ inter-item correlations. Frequencies, correlations, and χ tests for differences were used for item-level analyses.

To evaluate the VA single-item measure and different scoring options of the PHQ, we used bivariate analyses (correlation and *t* tests) and ROC analysis. Through ROC analysis, the sensitivity and specificity of the study measure are assessed using a more established measure of disease status as the reference standard. The area under the curve (AUC) can range from 0 to 1.0; an AUC of .50 suggests that classification based on the instrument under study is no more accurate than random chance. Analyses were performed using SPSS" version 11.5 for Windows (SPSS Inc, Chicago, Ill); a Web-based clinical calculator was used to calculate likelihood ratios.[31]

RESULTS

Internal consistency for the PHQ-9 was excellent (α = .86).[32] No item detracted from the consistency of the scale; inter-item correlations ranged from .27 to .68. The strongest associations were between anhedonia and depressed mood (r = .68) and self-esteem and depressed mood (r = .62); the psychomotor and self-harm items had slightly weaker inter-item associations (ranging from .27 to .44) than the other items.

The distribution of PHQ-9 scores (n = 1211) was positively skewed (mean = 4.76, median = 2, SD = 6.16). Using the cut scores for depression severity, 436 (36.0%) patients had scores indicating at least mild depressive symptoms (PHQ-9 ≥ 5); 251 (20.7%), at least moderate depressive symptoms (PHQ-9 ≥ 10); 120 (9.9%), at least moderately severe depressive symptoms (PHQ-9 ≥ 15); and 39 (3.2%), severe depressive symptoms (PHQ-9 ≥ 20). Using the major depression algorithm based on DSM-IV criteria, 12% of patients met the criteria for a provisional diagnosis of depression.

Of the 1211 study patients, 973 (80.3%) responded "no" to the VA single-item depression screen and 238 (19.7%) responded "yes" (**Table 1**). Table 1 presents single-item depression screen results by depression severity based on the PHQ-9. Patients with positive single-item screens had significantly higher PHQ-9 scores than those with negative screens (12.90 vs 2.77, t = 23.36, P < .001). Nearly 9 out of 10 patients (89.5%) with a positive single-item screen had PHQ scores indicating at least mild symptoms of depression. On the other hand, 8.2% of patients with PHQ-9 scores suggesting moderate to severe depression did not endorse the single-item screen.

Table 2 presents sensitivities, specificities, likelihood ratios, and areas under the ROC curve (AUCs) for the VA single-item screen and the PHQ-2, with the 3 main scoring algorithms of the PHQ-9 as reference standards. For each standard, PHQ-2 ≥ 2 demonstrated greater sensitivity than the VA single-item screen. PHQ-2 ≥ 3 is more sensitive than the VA-single item when using the major depression algorithm as the reference standard, but the confidence intervals slightly overlap at PHQ-9 ≥ 10 and = 15. In turn, PHQ-2 ≥ 3 is as sensitive as PHQ-2 = 2 (ie, the confidence intervals overlap), except when screening for moderate depression symptoms (PHQ-9 ≥ 10). In terms of specificity, the VA single-item outperformed PHQ-2 ≥ 2 but not PHQ-2 ≥ 3, for which the differences are not statistically significant. Finally, the AUC for PHQ-2 ≥ 2 was greater than the AUC for the VA single-item screen when using PHQ-9 ≥ 10 and the major depression algorithm; the AUC for PHQ-2 ≥ 3 was greater than the AUC for the VA single-item screen when using the more stringent standards (PHQ-9 ≥ 15 and the major depression algorithm).

Eighty (6.6%) patients rated the item "thoughts that you would be better off dead or of hurting yourself" as occurring at least several days over the past 2 weeks. In response to the follow-up questions, 28 (2.3%) acknowledged thoughts of harming themselves and 16 (1.3%) acknowledged having a specific plan. **Table 3** displays the percentage of patients with potential sui-

cidal ideation who would have been identified by the VA single-item screen and by PHQ scores using different PHQ formats and scoring methods. Of note, the correlation between PHQ-9 scores and PHQ-8 (which excludes the death/suicide item) scores was very high (r = .998, P < .001, n = 1044); only 3 patients with a PHQ-9 score of 10 or higher had a PHQ-8 score of less than 10.

Table 1. Distribution of PHQ-9 Scores by Response to VA Single-Item Screen in 1211 VA Participants*

	Responses to VA Single-Item Screen	
PHQ-9 Score: Depression Severity	Negative	Positive
0-4: Not depressed	750 (77.1%)	25 (10.5%)
5-9: Mild	143 (14.7%)	42 (17.6%)
10-14: Moderate	58 (6.0%)	73 (30.7%)
15-19: Moderately severe	16 (1.6%)	65 (27.3%)
20-27: Severe	6 (0.6%)	33 (13.9%)
Total	973	238

*PHQ-9 indicates the 9-Item Patient Health Questionnaire.

DISCUSSION

The data presented here suggest that while the VA single-item depression screen is specific, it is only moderately sensitive when the PHQ-9 cut point for moderate depression is used as the reference standard. Although changing the definition of a "positive" PHQ-9 score from ≥10 to ≥15 brings the single item's sensitivity to within the range of those recorded for case-finding measures,[12] raising the bar so that only those with moderately severe depression or major depression are detected may be inappropriate for screening conducted in a primary care setting.[11,13] In comparison to the VA single-item screen, the PHQ-2 performed very well. Using a PHQ-2 cut point of ≥ 2 rather than ≥ 3 improves its sensitivity, but also increased the false-positive rate; using the PHQ major depression algorithm as a reference standard,

Table 2. Performance of the VA Single-Item Screen* and PHQ-2 using Alternative PHQ Scoring Methods as Reference Standards†

Screen	% Sensitivity (95% CI)	% Specificity (95% CI)	+LR	AUC (95% CI)
PHQ ≥ 10 as reference standard				
VA single item	68 (62, 74)	93 (91, 94)	9.73	.81 (.77, .84)
PHQ-2 ≥ 2	95 (91, 97)	89 (87, 91)	8.43	.92 (.90, .94)
PHQ-2 ≥ 3	76 (71, 81)	95 (94, 97)	16.69	.86 (.83, .89)
PHQ ≥ 15 as reference standard				
VA single item	82 (74, 88)	87 (85, 89)	6.36	.84 (.80, .89)
PHQ-2 ≥ 2	99 (95, 100)	79 (77, 82)	4.77	.89 (.87, .91)
PHQ-2 ≥ 3	93 (87, 97)	89 (86, 90)	8.21	.91 (.88, .94)
Major depression algorithm§				
VA single item	78 (71, 84)	88 (86, 90)	6.77	.83 (.79, .87)
PHQ-2 ≥ 2‖	100 (97, 100)	81 (79, 84)	5.35	.91 (.89, .92)
PHQ-2 ≥ 3	97 (93, 99)	91 (89, 93)	11.13	.94 (.92, .96)

*Have you felt depressed or sad most of the time in the past year?
†PHQ-2 is the first 2 items of the PHQ-9. ("Over the last 2 weeks, how often have you been bothered by… 1. Little interest or pleasure doing things; 2. Feeling down, depressed or hopeless?")
‡AUC indicates area under the receiver operating characteristic curve; CI, confidence interval; LR+, positive likelihood ratio; PHQ, Patient Health Questionnaire. Positive likelihood ratio = sensitivity/(1 - specificity).
§Major depression algorithm = 5 or more symptoms (must include anhedonia or depressed mood) more than half the days; suicide ideation counts regardless of frequency.[17]
‖By definition, a score of less than 2 on the PHQ-2 is classified as "not depressed" using the major depression algorithm, so at a cut point of 2, the PHQ-2 should not produce false-negatives.

58% of those in our study who scored positive at PHQ-2 ≥ 2 were false positives. Whichever cut point is selected, if a very short screen is to be used, our results suggest that the PHQ-2 surpasses the single-item screen in a VA primary care setting, particularly in terms of sensitivity.

Notably, our findings for a single-item depression measure differ from those of Williams et al[26] (78% vs 85% sensitive and 88% vs 66% specific, respectively). The higher sensitivity Williams et al report may stem from their use of a clinician interview as the reference standard, differences in the sample patients' sex and ethnicity, or, less likely, the slight difference in wording. In contrast, our findings for the PHQ-2 are more consistent with those of Kroenke et al,[25] who studied 580 patients (66% women, 21% ethnic minority, mean age of 46 years) from community-based primary care and obstetrics-gynecology clinics. Using PHQ-2 ≥ 3, Kroenke et al reported 83% sensitivity, 92% specificity, and an AUC of .93, again with a structured interview by a mental health clinician as the reference standard. We found 97% sensitivity, 91% specificity, and an AUC of .94 with the PHQ major depression algorithm as a reference standard. These results suggest that the 2-item screen is more generalizable across patient populations than the 1-item screen and/or that the impact of sample demographics on brief screen performance may be substantial. Indeed, Williams et al report predictive differences by ethnicity, and Kroenke et al found that patient age (but not patient sex) affected the results.

In our sample, scores based on the PHQ-8 signaled depression in all but 1 patient who expressed an active suicide plan. That is, administering the final PHQ-9 item assessing suicidal ideation did not improve case-finding over the PHQ-8. Importantly, however, approximately one third of the patients who endorsed the PHQ-9 death or suicide item in our study had active suicidal ideation and received urgent clinical attention, which would not have occurred had they not been administered the item addressing thoughts of death or self-harm. Thus, for clinical purposes we recommend the following algorithm: if a patient responds affirmatively to either of the PHQ-2 items, the remaining 7 items of the PHQ-9 should be administered. If the PHQ-9 score suggests major depression or suicidal ideation, clinicians must be prepared to conduct further assessment and to offer or arrange for appropriate treatment.

Our data suggest that 1 of every 3 veterans seen in primary care who is not already receiving specialty treatment has symptoms consistent with at least mild depression, 1 in 5 has symptoms consistent with at least moderate depression, and 1 in 10 has symptoms consistent with moderately severe depression. A prevalence of approximately 20% is congruent with previous estimates from veteran samples.13,34 Yet our prevalence estimate is alarming in that, in contrast to previous VA studies,[13,33] we excluded patients who had seen a mental health professional in the last 6 months.

It is important to note several limitations of this study. First, we used the PHQ and not a formal diagnostic interview as our reference standard. The PHQ, however, has strong, well-documented psychometric properties, and our PHQ-2 results are comparable to those of Kroenke et

Table 3. Detecting Depression in Patients With Potential Suicide Ideation Using the VA Single-Item Screen and Alternative PHQ-9 Scoring Methods*

		Detected by…					
Item Endorsed	**No.**	**VA Single-Item Screen**	**PHQ-9 ≥ 10**	**PHQ-8 ≥ 10**	**PHQ-2 ≥ 3**	**PHQ-2 ≥ 2**	**Major Depression Algorithm**
Thoughts that you would be better off dead or hurting yourself in any way†	80	58 (72.5%)	71 (88.8%)	68 (85.0%)	59 (73.8%)	74 (92.5%)	54 (67.5%)
Thoughts of hurting or killing yourself (vs thoughts that you would be better off dead)‡	28	23 (82.1%)	26 (92.9%)	25 (89.3%)	22 (78.6%)	26 (92.9%)	19 (67.9%)
Recent thoughts of any specific ways to hurt or kill yourself§	16	11 (68.8%)	15 (93.8%)	15 (93.8%)	13 (81.3%)	15 (93.8%)	10 (62.5%)

*PHQ indicates Patient Health Questionnaire.
†Item 9 of the PHQ-9; endorsement = any response > 0 (> never).
‡The follow-up item that was administered to patients endorsing item 9 of the PHQ-9.
§The follow-up item that was administered to patients acknowledging thoughts of hurting or killing oneself. Follow-up items use yes/no response format.

al,[25] who used a mental health clinician interview as a reference standard. Second, the generalizability of our results to nonveteran populations may be limited. Individuals without telephones, who were severely depressed, or who otherwise were unable to complete a phone screening might be underrepresented. Third, the associations we detected may be inflated by various factors. For instance, all PHQ-9 items are keyed in the same direction, which can contribute to response sets and exaggerate internal consistency. Also, the PHQ-2 is drawn directly from the PHQ-9. However, scores based on the first 2 items of the PHQ correlate with scores based on the last 7 items almost as strongly as with PHQ-9 scores (r = .76 vs r = .88), suggesting that the association between the PHQ-2 and the PHQ-9 is not solely an artifact of common items. Finally, we administered the PHQ-9 and the single-item screen concurrently, although the influence of this is likely to be slight.[34]

If the goal of screening is to identify potential cases of depression, then the following must be considered. Current guidelines suggest that all general medicine patients not already being seen by mental health professionals should be screened, with repeat screening if risk factors or symptoms are present and systems are in place to support diagnosis and treatment.[4,6] In depression screening instruments, sensitivity is critical.[11,13] Thus, the single VA item is less than optimal. Administering 2 items improves performance with minimal added time investment, and using PHQ-2 ≥ 2 results in appropriate levels of sensitivity.

Although brief instruments can facilitate screening programs in primary care settings, these instruments are not sufficient to confirm the diagnosis of depression or the severity of suicidal ideation. Clinician assessment must follow. In making a diagnosis, the clinician should take into account the patient's history, comorbidities, functional status, and safety —considerations precluded in any brief screening instrument.

Acknowledgments

We wish to thank Nancy Cuilwik, BS, LeAnn Snodgrass, Megan Crutchfield, BS, and Jeff Solodky, BA, for their help in data organization and analysis, and manuscript preparation.

REFERENCES

1. Kessler RC, Berglund P, Demler O, et al. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA*. 2003;289:3095-105.

2. Simon GE, VonKorff M. Recognition, management, and outcomes of depression in primary care. *Arch Fam Med*. 1995;4:99-105.

3. Lebowitz BD, Pearson JL, Schneider LS, et al. Diagnosis and treatment of depression in late life: consensus statement update. *JAMA*. 1997;278:1186-1190.

4. *Management of Major Depressive Disorder in Adults in the Primary Care Setting.* Washington, DC: VA/DoD Evidence Based Clinical Practice Guideline Working Group, Veterans Health Administration, Department of Veterans Affairs, and Health Affairs, Department of Defense; May 2000. Office of Quality and Performance publication 10Q-CPG/MDD-00. Available at: http://www.oqp.med.va.gov/cpg/MDD/MDD_Base.htm. Accessed March 26, 2004.

5. *U.S. Preventative Services Task Force Guide to Clinical Preventive Services.* Washington, DC: US Department of Health and Human Services, Office of Disease Prevention and Health Promotion; 1996.

6. U.S. Preventative Services Task Force. Screening for depression: recommendations and rationale. *Ann Intern Med*. 2002;136:760-764.

7. Williams JW, Hitchcock NP, Cordes JA, Ramirez G, Pignone M. Rational clinical examination: is this patient clinically depressed? *JAMA*. 2002;287:1160-1167.

8. VHA/DoD Guideline for Major Depressive Disorder: Module A. FY2002 VHA Performance Measurement System: Technical Manual. March 8, 2002:58-61. Washington, DC: Office of Quality and Performance; 2002. Available at: http://www.oqp.med.va.gov/cpg/MDD/P/MDD_3_8_02_techman.doc. Accessed September 15, 2003.

9. Nutting PA, Rost K, Dickinson M, et al. Barriers to initiating depression treatment in primary care practice. *J Gen Intern Med*. 2002;17:103-111.

10. Rost K, Nutting P, Smith J, et al. The role of competing demands in the treatment provided primary care patients with major depression. *Arch Fam Med*. 2000;9:150-154.

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

12. Williams JW, Pignone M, Ramirez G, Perez Stellato C. Identifying depression in primary care: a literature synthesis of case-finding instruments. *Gen Hosp Psychiatry*. 2002;24:225-237.

13. Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression: two items are as good as many. *J Gen Intern Med*. 1997;12:439-445.

14. Chochinov HM, Wilson KG, Enns M, Lander S. "Are you depressed?" Screening for depression in the terminally ill. *Am J Psychiatry*. 1997;154:674-676.

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

16. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric Ann*. 2002;32:509-515.

17. Spitzer RL, Kroenke K, Williams JBW, et al. Validation and utility of a self-report version of PRIME-MD. *JAMA*.1999;282:1737-1744.

18. Dietrich AJ, Oxman TE, Burns MR, Winchell CW, Chin T. Application of a depression management office system in community practice: a demonstration. *J Am Board Fam Pract*. 2003;16(2):107-114.

19. Diez-Quevedo C, Rangil T, Sanchez-Planell L, Kroenke K, Spitzer RL. Validation and utility of the patient health questionnaire in diagnosing mental disorders in 1003 general hospital Spanish inpatients. *Psychosom Med*. 2001;63:679-686.

20. Ruo B, Rumsfeld JS, Hlatky MA, Liu H, Browner WS, Whooley MA. Depressive symptoms and health-related quality of life: The Heart and Soul Study. *JAMA*. 2003;290:215-221.

21. Löwe B, Spitzer RL, Gräfe K, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disord*. 2004;78:131-140.

22. Löwe B, Gräfe K, Zipfel S, Witte S, Loerch B, Herzog W. Diagnosing ICD-10 depressive episodes: superior criterion validity of the Patient Health Questionnaire. *Psychother Psychosom*. 2004;73:386-390.

23. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the PHQ-9. *Med Care*. In press.

24. Kroenke K, Unützer J, Callahan CM, et al. Monitoring depression with a brief self-report scale (PHQ-9) [abstract]. *J Gen Intern Med*. 2004[Sl];19:181.

25. Kroenke K, Spitzer RL, Williams JBW. The patient health questionnaire-2: validity of a two-item depression screener. *Med Care*. 2003;41:1284-1292.

26. Williams JW, Mulrow CD, Kroenke K et al. Case-finding for depression in primary care: a randomized trial. *Am J Med*. 1999;106:36-43.

27. Richardson C, Waldrop J. Census 2003 Brief. US Department of Commerce, Economics and Statistics Administration, US Census Bureau; 2003. US Census Bureau publication C2 KBR-22. Available at: http://www.va.gov/vetdata/Census2000/c2kbr-22.pdf . Accessed April 9, 2004.

28. Wells KB, Burnam MA, Leake B, et al. Agreement between face-to-face and telephone administered versions of the depression section of the NIMH Diagnostic Interview Schedule. *J Psychiatr Res*. 1988;22:207-220.

29. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. Washington, DC: American Psychiatric Association;1994.

30. Roth PL. Missing data: a conceptual review for applied psychologists. *Personnel Psychol*. 1994;47:537-560.

31. Division of General Internal Medicine, Medical School of Wisconsin. Online clinical calculators. Available at: http://www.intmed.mcw.edu/clincalc.html. Accessed September 24, 2003.

32. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic*. 1981;86:127-137.

33. *Health Status and Outcomes of Veterans: Physical and Mental Component Summary Scores, Veterans SF-36, 1999 Large Health Survey of Veteran Enrollees.* Washington, DC: Department of Veterans Affairs, Veterans Health Administration, Office of Quality and Performance; 2000.

34. Rost K, Burnam MA, Smith GR. Development of screeners for depressive disorders and substance disorder history. *Med Care*. 1993;31:189-200.