

RESEARCH ARTICLE

Open Access

A study of the diagnostic accuracy of the PHQ-9 in primary care elderly

Elizabeth Phelan^{1,2*}, Barbara Williams², Kathryn Meeker¹, Katie Bonn³, John Frederick³, James LoGerfo^{2,4}, Mark Snowden³

Abstract

Background: The diagnostic accuracy of the Patient Health Questionnaire-9 (PHQ-9) for assessment of depression in elderly persons in primary care settings in the United States has not been previously addressed. Thus, the purpose of this study was to evaluate the test performance of the PHQ-9 for detecting major and minor depression in elderly patients in primary care.

Methods: A prospective study of diagnostic accuracy was conducted in two primary care, university-based clinics in the Pacific Northwest of the United States. Seventy-one patients aged 65 years or older participated; all completed the PHQ-9 and the 15-item Geriatric Depression Scale (GDS) and underwent the Structured Clinical Interview for Depression (SCID). Sensitivity, specificity, area under the receiver operating characteristic (ROC) curve, and likelihood ratios (LRs) were calculated for the PHQ-9, the PHQ-2, and the 15-item GDS for major depression alone and the combination of major plus minor depression.

Results: Two thirds of participants were female, with a mean age of 78 and two chronic health conditions. Twelve percent met SCID criteria for major depression and 13% minor depression. The PHQ-9 had an area under the curve (AUC) of 0.87 (95% confidence interval [CI], 0.74-1.00) for major depression, while the PHQ-2 and the 15-item GDS each had an AUC of 0.81 (95% CI for PHQ-2, 0.64-0.98, and for 15-item GDS, 0.70-0.91; $P = 0.551$). For major and minor depression combined, the AUC for the PHQ-9 was 0.85 (95% CI, 0.73-0.96), for the PHQ-2, 0.80 (95% CI, 0.68-0.93), and for the 15-item GDS, 0.71 (95% CI, 0.55-0.87; $P = 0.187$).

Conclusions: Based on AUC values, the PHQ-9 performs comparably to the PHQ-2 and the 15-item GDS in identifying depression among primary care elderly.

Background

The incidence of major depression in the general elderly population is approximately 15% per year and doubles after age 70 [1]. The prevalence of depression is higher in medical settings than in the community. Depression is associated with an increased risk of mortality [2], healthcare utilization [3], functional decline [4-6], and poorer quality of life [7,8]. Depression is however, quite responsive to treatment, and antidepressants are as effective for older adults as for younger individuals [1]. Care management has been shown to improve outcomes for elderly who are depressed and treated in primary care settings [9]. Thus, detection of depression among

older adults in primary care is important, as it can be linked to effective treatment. Additionally, because most older adults seek care for their mental health issues in primary care [10], screening in primary care with referral to a mental health specialist for diagnostic evaluation of depressive symptoms is neither efficient nor practical in most instances.

To date, late life depression has been under-recognized and inadequately treated [11-13]. A fundamental challenge for the primary care provider is recognition of depression in the elder where depression symptoms and comorbid medical illness symptoms such as insomnia or anorexia overlap [14,15]. This situation is further complicated by frequent underreporting of depressive symptoms by older adults [12]. A brief screening tool that accurately identifies depression among elderly patients would make the identification of depression in primary

* Correspondence: phelane@u.washington.edu

¹Department of Medicine, Division of Gerontology and Geriatric Medicine, University of Washington, Seattle, WA, USA

Full list of author information is available at the end of the article

care more straightforward, and improvements in identification and treatment of depression in elders might lead to improved function, survival, and quality of life. Thus, our study objective was to assess the diagnostic performance of the Patient Health Questionnaire-9 (PHQ-9) [16], a promising depression scale that has been validated with younger populations, and compare its performance to an established depression screening instrument, the 15-item Geriatric Depression Scale (GDS) [17], among elderly persons receiving health care in primary care settings. The PHQ-9 is the first self-report questionnaire designed for use in primary care that reflects the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) diagnostic criteria for depression, and so (through examining the pattern and number of items endorsed) can be used as a diagnostic tool for major and minor depression [16]. By contrast, other self-report instruments, including the GDS, do not map to the DSM depression diagnostic criteria and so cannot be used for depression diagnosis (only screening) [10], meaning that additional evaluation to establish a depression diagnosis must be conducted following a positive screen. While the PHQ-9 has been shown to be useful in general populations [16,18], including outpatient clinic settings in a variety of countries [19,20] and sensitive to change in the elderly [21], to our knowledge, its screening characteristics (sensitivity and specificity) had not been validated with elderly in a primary care setting in the United States. If a screening instrument's summary score (obtained through summing individual items scores) correlated strongly with gold-standard depression diagnosis obtained via in-person diagnostic interviews, this would be attractive, as it would make depression detection in primary care much more straightforward. A validation study examining this question is important, as the PHQ is being administered to adults of all ages, including the elderly, as part of national surveillance efforts to estimate the prevalence of mental health disorders in the United States [22]. We located only one study that focused specifically on the validity of the PHQ-9 in chronically ill elderly in primary care; this study was conducted in the Netherlands and focused on elders with diabetics and chronic obstructive pulmonary disease without known depression enrolled in a randomized trial of a nurse-led, psychological intervention [23]. We hypothesized that, because the PHQ-9 items map directly to DSM-IV criteria for depression, the PHQ-9 would have test performance characteristics at least comparable to the GDS but perhaps less ease of use (due to its more complex response format). We were interested in minor as well as major depression, because minor depression is more prevalent in primary care than major depression, is associated with adverse

effects on functioning and may increase the risk of subsequent major depression in older primary care patients [24,25], increases health care use and costs [24], and may be responsive to treatment [26-28]. Because brief screening tools may be preferable for use in primary care settings, we also examined the sensitivity and specificity of the PHQ-2, an abbreviated version of the PHQ-9, in identifying patients with major and minor depression.

Methods

Setting

The study was conducted from November 2006 through August 2007 in two primary care clinics affiliated with the University of Washington, in Seattle, Washington, USA. These clinics were chosen because they provide primary care for elders and because clinic administrators and providers were supportive of the project.

Study sample and procedures

Participants were consecutive established patients aged 65 or older presenting to the clinics for care. Those for whom the study procedures were not feasible due to severe dementia, unstable medical condition, or non-English fluency, were excluded. After informed consent was obtained, the PHQ-9 and 15-item GDS were administered to each participant by a research assistant. A geriatric psychiatrist or gerontologic psychiatric nurse practitioner trained in SCID administration, blinded to the results of the depression screening tests, conducted a diagnostic interview for depression, the Structured Clinical Interview for Depression (SCID). The SCID is considered the criterion standard for DSM-IV depression diagnosis in clinical research [29]. Information on demographic characteristics (age, gender, race) and chronic conditions was obtained from a questionnaire completed by the participant at the time of enrollment. All instruments were administered on the same day in the context of a routine clinic appointment. To avoid ordering effects, the order of administration of screening instruments and the SCID was varied in random fashion. The institutional review board at the University of Washington approved all procedures.

Depression instruments

The PHQ-9 is a self-administered, nine-item questionnaire specific to depression that is available free to end users [16]. It was developed as a self-report version of the Primary Care Evaluation of Mental Disorders (PRIME-MD). It has several other features that make it attractive for use with older adults in primary care settings, including being substantially shorter than most other depression screening measures; having been originally developed and tested for use with medical

patients, who are likely (as are elderly) to have high rates of physical symptoms consistent with either chronic medical illness or depression; having excellent test-retest reliability, excellent criterion [16] and construct validity [18], and responsiveness [21,30], or the capacity of an instrument to detect meaningful change over time [31].

Each of the nine items of the PHQ-9 is scored as 0 (not at all), 1 (several days), 2 (more than half the days), or 3 (nearly every day). As a screening tool, summing the 9 items, (score range 0-27 with 0 indicating no depressive symptoms and 27 indicating all symptoms occurring nearly daily), a score of ≥ 10 has been shown to have an 88% sensitivity and 88% specificity for major depression in a general medical population [18].

As a diagnostic tool, major depression is diagnosed if ≥ 5 of the 9 symptoms elicited have been present at least more than half the days in the past two weeks AND one of these symptoms is either depressed mood or anhedonia [18]. Minor depression is diagnosed if 2-4 symptoms have been present at least more than half the days in the past two weeks AND one of the symptoms is either depressed mood or anhedonia [18].

Depression severity can also be assessed with the PHQ-9 [16]. Kroenke et al suggested cutpoints to identify minimal (0-4), mild (5-9), moderate (10-14), moderately severe (15-19), and severe (≥ 20) depression [18]. These cutpoints have some empiric support, as demonstrated by a strong association between increasing PHQ-9 scores and worse scores on functional status (SF-20) measures, especially for scales most strongly correlated with depression [18].

The PHQ-2 is an abbreviated version (i.e., the first two items) of the PHQ-9 that inquires about depressed mood and anhedonia; it has been studied to a lesser extent than the PHQ-9 as a depression screening tool [32-34]. In studies of younger adults in primary care and obstetrics-gynecology settings, scores of ≥ 3 have a sensitivity of 83% and a specificity of 92% for major depression [33].

The 15-item GDS is the most commonly used, geriatric specific depression screen and thus represents the current "usual care" standard for geriatric depression screening [17]. Its brevity and dichotomized (yes, no) response format offer some ease of use advantages. In a study of persons 60 years or older from primary care practices, scores of > 5 had a sensitivity of 92% and a specificity of 81% for major depression [17].

Published cutpoints for minor depression are not available for the three instruments for purposes of screening. However, the PHQ-9 can be used as a diagnostic tool for minor depression, since the items map directly to DSM minor depression criteria.

Statistical analysis

Our analysis was conducted in three stages. First, descriptive statistics were calculated to characterize the sociodemographic and health characteristics of our sample. Next, sensitivity, specificity, and likelihood ratios (LRs) for detecting major depression for each of the three instruments (i.e., including the PHQ-2) were calculated over a range of cutpoints. Ninety-five percent confidence intervals (95% CI) were calculated for sensitivity and specificity using an online clinical calculator, available at <http://statpages.org/ctab2x2.html>. Third, receiver operating characteristic (ROC) analyses were conducted. ROC analyses combine instrument sensitivity and specificity into one measure (referred to as area under the curve, or AUC) for all possible cutpoints. AUC values range from ≤ 0.5 (no discriminatory ability) to 1.0 (perfect discrimination). An AUC of 0.84 implies that there is an 84% likelihood that a randomly selected person with depression will have a higher PHQ-9 score than a randomly selected non-depressed person. The AUC was measured to permit comparison of the diagnostic value of each instrument for detecting major depression, for the study group overall and by gender, ethnicity, age, and burden of comorbidity. A global non-parametric test for comparison of the 3 AUCs was calculated [35]. Lastly, we calculated sensitivity, specificity, LRs, and AUCs with depression broadened to include minor depression. Statistical analyses were performed using STATA v. 9.2 (Stata Corp., College Station, Texas).

Results

Participant flow

A total of 502 unique, established patients were seen in the clinics during the study period. Of these, 122 were not approached due either to being non-fluent in English ($N = 64$) or having severe dementia ($N = 58$). Of the 380 remaining, 227 met the age criterion and so were approached about the study by clinic staff. Of these, 121 were willing to speak with the research assistant about the study, and 71 agreed to participate.

Baseline characteristics of participants

The 71 participants had a mean age of 78 years; nearly two-thirds were female, a third were non-white, and over three-quarters had a high school education (Table 1). They reported having two chronic medical conditions, on average, with hypertension, arthritis, and diabetes being the most common (reported by 63%, 49%, and 24%, respectively). About a quarter had a PHQ-9 score of 10 or greater, and one-fifth had a PHQ-2 score of three or greater. About half had a GDS score greater than five. The SCID was positive for major depression

Table 1 Demographic and health characteristics of participants at enrollment (N = 71)

Characteristic	
Age, years, mean ± SD*	78 ± 7
Female, %	62
Non-white, %	32
High school graduate, %	82
Chronic medical conditions	
Mean ± SD	2.3 ± 1.5
Median (interquartile range)	2 (2)
Three or more chronic medical conditions, %	42
Patient Health Questionnaire-9 (PHQ-9)	
Score, mean ± SD	5.9 ± 6.1
Score ≥10, % [†]	23 (16/71)
Patient Health Questionnaire-2 (PHQ-2)	
Score, mean ± SD	1.2 ± 1.6
Score ≥3, % [‡]	20 (14/71)
15-Item Geriatric Depression Scale (GDS)	
Score, mean ± SD	5.8 ± 1.8
Score >5, % [§]	48 (33/69)
Structured Clinical Interview for Depression (SCID)	
Major depression, %	12 (8/69)
Minor depression, %	13 (9/69)

* SD = standard deviation.

[†] Scores of 10 or greater on the PHQ-9 suggest likely major depression in general medical populations [18].

[‡] Scores of 3 or greater on the PHQ-2 suggest possible depression in adults in primary care and obstetrics-gynecology settings [33].

[§] Scores greater than 5 on the 15-item GDS suggest major depression in older adults in primary care [17].

in 12% and for minor depression in 13%. Thirty percent needed help to complete the demographics questionnaire, 30% needed help to complete the GDS, and 37% needed help to complete the PHQ-9. Assistance was usually in the form of oral administration of the measures due to poor eyesight or difficulty using a pen. The yes/no response option format of the GDS was not clearly easier for participants to use than the four option response format of the PHQ-9. Participants took about five minutes to complete the PHQ-9, and about 25-50% longer to complete the GDS, on average.

Relative performance of depression screening instruments for major depression

The sensitivity and specificity of the screening measures were calculated using the SCID as the criterion standard for major and minor depression diagnosis. Using published, standard major depression cutpoints for these tests, seven of the eight (63%) majorly depressed participants were correctly classified with the PHQ-9 (cutpoint ≥ 10) and the PHQ-2 (cutpoint ≥ 3), while 100% were correctly classified with the GDS (cutpoint > 5) (Table 2). At these published cutpoints, the specificity was higher for the two PHQ measures (PHQ-9, 82%; PHQ-

2, 85%) compared to the 15-item GDS, with a specificity of 58%. The LR positive represents how much the odds of having depression increase when a test is positive. At the standard cutpoints, the two PHQ measures had LR positives of about four, – i.e., a positive screen is four times more likely to be seen in someone with major depression than in someone without – while the 15-item GDS had a LR positive of 2.4, – i.e., a positive screen is 2.4 times more likely to be seen in someone with major depression than in someone without the condition.

Table 3 reports results from the ROC analysis, which gives a global assessment of the discriminatory power of each instrument. Overall, the AUC for the PHQ-9 was comparable to the AUC for the PHQ-2 and the 15-item GDS (0.87 for PHQ-9 vs. 0.81 for both the PHQ-2 and the 15-item GDS, $P = 0.551$).

As shown in Table 3, the AUC for the PHQ-9 and 15-item GDS was similar for men and women, whereas for the PHQ-2, the AUC was lower for women. This same pattern of AUCs for the three instruments held for white and non-white participants. The PHQ-9 appeared somewhat more discriminatory for those under the age of 80 years, whereas the PHQ-2 and the 15-item GDS discriminated comparably for these subgroups. The PHQ-9 and PHQ-2 AUC values were highest for participants with less than three comorbidities (0.93 and 0.92, respectively), whereas the AUC for the 15-item GDS was similar regardless of comorbidity burden.

Relative performance of depression screening instruments using broadened definition of depression

When the broadened definition of depression was used (i.e., including minor and major depression), the sensitivity of all three measures worsened as compared with their case detection for major depression alone (Table 4). For example, at the PHQ-9 cutpoint of ≥10, the sensitivity using the broadened definition was 59% as opposed to 63%; at the PHQ-2 cutpoint of ≥3, the sensitivity was 53% as opposed to 63%. For the 15-item GDS, at the cutpoint of >5, the sensitivity was 81% as opposed to 100%. AUC values using the broadened definition of depression were: PHQ-9 = 0.85 (95% CI, 0.73-0.96); PHQ-2 = 0.80 (95% CI, 0.68-0.93); and 15-item GDS = 0.71 (95% CI, 0.55-0.87); $P = 0.187$ for comparison of AUCs for the three instruments.

Discussion

This study demonstrated that, used as a screening instrument, the PHQ-9 performed comparably to the PHQ-2 and the 15-item GDS for the purposes of major depression detection in elderly individuals in primary care. Broadening the definition of depression to include minor along with major depression did not improve

Table 2 Sensitivity and specificity of depression screening instruments for diagnosing major depression at various cutpoints*

Instrument and Cutpoint	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	+ Likelihood Ratio	- Likelihood Ratio
PHQ-9 Cutpoint				
≥8	88 (56-98)	75 (71-77)	3.6	0.16
≥9	88 (56-98)	80 (76-82)	4.4	0.16
≥10	63 (33-86)	82 (78-85)	3.5	0.46
≥11	63 (33-85)	84 (80-87)	3.8	0.45
≥12	63 (33-85)	84 (80-87)	3.8	0.45
PHQ-2 Cutpoint				
≥1	88 (55-98)	61 (56-62)	2.2	0.21
≥2	75 (43-93)	67 (63-70)	2.3	0.37
≥3	63 (33-85)	85 (81-88)	4.2	0.44
≥4	38 (15-62)	93 (91-97)	5.7	0.67
≥5	38 (16-48)	98 (96-100)	22.9	0.64
15-item GDS Cutpoint				
≥5	100 (73-100)	15 (11-15)	1.2	0.00
≥6	100 (70-100)	58 (54-59)	2.4	0.00
≥7	75 (43-93)	77 (72-79)	3.2	0.33
≥8	25 (7-54)	87 (84-91)	1.9	0.86
≥9	13 (2-37)	93 (92-97)	1.9	0.94

* Cutpoints are specific sum scores that distinguish between individuals with and without the disorder. Bolded cutpoints indicate the optimum balance between sensitivity and specificity, while italicized cutpoints are those that are typically cited in the literature as being those that optimize sensitivity and specificity for the detection of major depression for each respective instrument when applied to the general population.

† CI = confidence interval.

performance of any of the three screening instruments. Self-administration of the PHQ-9 did not result in substantially greater need for assistance, compared to the GDS, among our study sample. The PHQ-9 performed comparably regardless of gender or race and somewhat better for younger elders and for those with less chronic illness. The prevalence of major depression in our sample was 12%, comparable to that found in other studies of elderly across a range of health care settings [36], and the prevalence of minor depression was 13%.

This study was motivated by a desire to address the challenge of depression recognition by non-psychiatric physicians, and in particular, recognition of depression in the elderly where depression symptoms and symptoms of comorbid medical illness overlap [13-15]. The utility of the PHQ-9 for detection of depression in other populations has been previously documented [16,18,37-46]; however, to our knowledge, ours is the first study to examine its screening characteristics for both major and any (major or minor) depression with

Table 3 Receiver operating curve analyses for major depression for each screening instrument

	N*	PHQ-9 AUC [†] (95% CI) [‡]	PHQ-2 AUC (95% CI)	GDS AUC (95% CI)
Overall	69	0.87 (0.74-1.00)	0.81 (0.65-0.98)	0.81 (0.70-0.91)
Gender				
Male	26	0.88 (0.72-1.00)	0.91 (0.78-1.00)	0.76 (0.58-0.93)
Female	43	0.85 (0.61-1.00)	0.70 (0.37-1.00)	0.84 (0.70-0.98)
Race				
White	47	0.88 (0.73-1.00)	0.87 (0.75-0.99)	0.85 (0.74-0.95)
Non-white [§]	22	0.93	0.31	0.90
Age				
< 80 years	38	0.92 (0.81-1.00)	0.81 (0.54-1.00)	0.82 (0.68-0.96)
≥80 years	31	0.80 (0.52-1.00)	0.83 (0.64-1.00)	0.80 (0.62-0.98)
Comorbidities				
< 3	41	0.93 (0.82-1.00)	0.92 (0.85-1.00)	0.79 (0.55-1.00)
≥3	28	0.80 (0.57-1.00)	0.71 (0.43-0.99)	0.82 (0.67-0.97)

* N is for PHQ-9 and PHQ-2; GDS is missing for one participant.

[†] AUC = area under the receiver operating characteristic curve. Values range from ≤0.5 (no discriminatory ability) to 1.0 (perfect discrimination – in this case, of depressed from non-depressed).

[‡] CI = confidence interval.

[§] Only one of the 22 non-white participants had a positive SCID, and therefore no confidence interval could be calculated.

an exclusively older adult study sample in a primary care clinic setting in the United States. With an AUC for major depression in our study of 0.87, the performance of the PHQ-9 in identifying major depression was somewhat worse than has been demonstrated for general medical populations, where the AUC was 0.95 [18]. In the only other published study that we were able to locate that focused specifically on the validity of the PHQ-9 in an elderly sample, the AUC for major depression was 0.92 [23]. It is of interest that, despite substantial differences between our two study samples (i.e., northern European vs. United States origin; enrolled in a randomized controlled trial vs. consecutively enrolled from a primary care clinic; excluded if carried a prior diagnosis of depression or other psychiatric disorder vs. depression and/or other psychiatric disorder other than severe dementia did not preclude participation) our results for major depression detection through use of the PHQ-9 are quite comparable. Taken together, they provide evidence in support of the instrument's validity as a screening instrument for elders with chronic illness, including those who carry a depression diagnosis or other psychiatric disorder.

Of note, the low specificity (i.e., 58%) of the 15-item GDS for major depression at the standard cutpoint (i.e., >5) may limit its use with older persons similar to those in

our study, because many would need follow-up evaluation to receive a specific depression diagnosis. By contrast, both the PHQ-9 and PHQ-2 have better sensitivity and specificity at the standard major depression cutpoints; however, compared to the 15-item GDS, both risk missing cases because of their lower sensitivity at those cutpoints.

It is of interest that the optimal cutpoints for detection of major depression for the instruments evaluated in this study varied from previously published cutpoints for general medical populations. Specifically, a PHQ-9 cutpoint of ≥9, a PHQ-2 cutpoint of ≥2, and a GDS cutpoint of ≥7 offered the best combination of sensitivity and specificity for our study participants. The need to modify cutpoints to achieve the best balance between sensitivity and specificity for elderly populations has been observed in other studies [23,36]. More research in this area appears warranted.

Even with modification of cutpoints, our results demonstrate that the PHQ-2 performs less well for detection of major depression among primary care elderly as compared to younger adults, where sensitivity and specificity have been reported to be 83% and 92% respectively [33]. Our finding is, however, consistent with the one study that examined the utility of the PHQ-2 for detecting major depression in elderly primary care patients and found a sensitivity of only 79% and specificity of 58% with a cutpoint of 1 or greater [36].

In contrast to other studies, wherein broadening the definition of depression to include depression of lesser severity and dysthymia improved the sensitivity of depression screening instruments [23,36], our study demonstrated that sensitivity was not improved through this maneuver. Sensitivity in our study was determined to have decreased as the result of an increase in false negatives that occurred when the definition of depression was broadened (data not shown). However, AUC values derived using our broadened definition remained comparable to those derived using the narrower (major depression only) definition and still in the acceptable range of discrimination.

This study was limited by several factors. First, we did not assess the percentage of patients already diagnosed with depression or the percentage on antidepressants. Second, our sample was in essence a convenience sample, rather than a nationally representative sample of elders receiving care in primary care settings. Additionally, our sample size was smaller than anticipated, due in part to substantial numbers of clinic patients being ineligible to participate because of language barriers (non-fluent in English) and cognitive dysfunction, and in part to a relatively low acceptance of study participation. The small sample size likely contributed to the small number (i.e., 8) of cases of major depression in the study, and this small number in turn limits interpretation of the statistical comparisons of the screening

Table 4 Sensitivity and specificity of each screening instrument using a broadened definition of depression*

Instrument and Cutpoint	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	+ Likelihood Ratio	- Likelihood Ratio
PHQ-9 Cutpoint				
≥6	77 (56-90)	69 (63-74)	2.5	0.34
≥7	77 (56-90)	77 (70-81)	3.3	0.31
≥8	77 (57-89)	83 (76-87)	4.4	0.28
≥9	71 (51-85)	87 (80-91)	5.2	0.34
≥10	59 (40-74)	89 (82-93)	5.1	0.47
PHQ-2 Cutpoint				
≥1	82 (62-94)	67 (61-71)	2.5	0.26
≥2	71 (50-86)	73 (66-78)	2.6	0.40
≥3	53 (35-67)	90 (85-95)	5.5	0.52
≥4	35 (21-40)	98 (94-100)	18.4	0.66
≥5	18 (7-22)	98 (95-100)	9.2	0.84
15-item GDS Cutpoint				
≥4	94 (85-99)	4 (1-5)	0.98	1.63
≥5	88 (71-96)	14 (8-16)	1.0	0.93
≥6	81 (60-93)	62 (55-65)	2.1	0.30
≥7	56 (36-74)	79 (73-84)	2.7	0.55
≥8	31 (16-46)	90 (86-95)	3.3	0.76

* Bolded cutpoints indicate the optimum balance between sensitivity and specificity.

* CI = confidence interval.

instruments. The low study participation may have affected the prevalence of depression in our study sample, as one could imagine that those with depression might be either more or less likely to agree to participate in a study of depression screening. Lastly, the validity of the depression instruments for identifying depression in persons with dementia or those who are non-English speaking could not be ascertained.

These limitations notwithstanding, this study has several strengths. First, our study was conducted with a diverse elderly sample, and we were able to describe the extent and nature of chronic medical conditions of study participants. Secondly, the depression screening instruments

were mostly self-administered, with assistance as needed from the research assistant, while the SCID was conducted by the study psychiatrist/study nurse practitioner, thus minimizing the likelihood of agreement of the screening instruments with the criterion standard. Thirdly, the psychiatrist and nurse practitioner who administered the SCID were blinded to the results of the depression screening instruments and to any treatment that the study participant may have already been receiving for depression.

Conclusions

In balance, these data suggest that the PHQ-9 performs comparably to the 15-item GDS when used as a screening

instrument for detection of depression among elderly persons in primary care settings. Because of its brevity and its utility in making specific, DSM-IV based, depression diagnoses, the PHQ-9 represents a reasonable alternative to the GDS, particularly in situations where referral to a mental health provider for definitive diagnostic evaluation is neither an option nor desired by the patient.

Acknowledgements

This research was funded by the small grants program of the Center for HealthCare improvement for Addictions, Mental Illness, and Medically vulnerable Populations (CHAMMP), Harborview Medical Center, University of Washington, Seattle, WA. The funder had no role in the study design, methods, data collection, analysis, or interpretation of data, nor any role in the preparation of the manuscript or decision to submit the manuscript for publication.

This research was conducted while Dr. Phelan was a K23 recipient from the National Institute on Aging and a Paul Beeson Physician Faculty Scholars in Aging Research Award recipient.

Author details

¹Department of Medicine, Division of Gerontology and Geriatric Medicine, University of Washington, Seattle, WA, USA. ²Department of Health Services, School of Public Health and Community Medicine, University of Washington, Seattle, WA, USA. ³Department of Psychiatry and Behavioral Sciences, Division of Geriatric Psychiatry, University of Washington, Seattle, WA, USA. ⁴Department of Medicine, Division of General Medicine, University of Washington, Seattle, WA, USA.

Authors' contributions

EAP conceptualized the design of the study, obtained funding for the project, worked closely with administrators and staff of participating clinics to ensure acceptance and understanding of the clinic's role in study procedures, obtained IRB approval for all study procedures, interpreted data, and had primary responsibility for preparation of the manuscript. BW performed data analyses and assisted with interpretation of the data and drafting of the manuscript. KM served as the research assistant. She recruited all study participants and ensured that they completed all study procedures. She was responsible for data tracking and management. KJB and JTF conducted diagnostic interviews for depression. JPL contributed to interpretation of the data, facilitated clinic involvement, and provided critical review of the manuscript. MS contributed to study conception and design, provided content expertise and logistical support for the project, interpreted data, and provided critical review and final approval of the manuscript. All authors approved the final manuscript.

Authors' information

Dr. Snowden is a practicing geriatric psychiatrist and director of the Geriatric Psychiatry Services Program at Harborview Medical Center, Seattle, Washington. He chaired an expert panel that conducted a systematic literature review of depression screening instruments for community-based (e.g., senior centers, adult day health, senior housing) settings.

Competing interests

The authors declare that they have no competing interests.

Received: 23 December 2009 Accepted: 1 September 2010
Published: 1 September 2010

References

- Alexopoulos GS: Depression in the elderly. *Lancet* 2005, **365**:1961-1970.
- Schulz R, Beach SR, Ives DG, Martire LM, Ariyo AA, Kop WJ: Association between depression and mortality in older adults: the Cardiovascular Health Study. *Arch Intern Med* 2000, **160**:1761-1768.
- Callahan CM, Hui SL, Nienaber NA, Musick BS, Tierney WM: Longitudinal study of depression and health services use among elderly primary care patients. *J Am Geriatr Soc* 1994, **42**:833-838.
- Stuck AE, Walthert JM, Nikolaus T, Bula CJ, Hohmann C, Beck JC: Risk factors for functional status decline in community-living elderly people: a systematic literature review. *Soc Sci Med* 1999, **48**:445-469.
- Sarkisian CA, Liu H, Gutierrez PR, Seeley DG, Cummings SR, Mangione CM: Modifiable risk factors predict functional decline among older women: a prospectively validated clinical prediction tool. The Study of Osteoporotic Fractures Research Group. *J Am Geriatr Soc* 2000, **48**:170-178.
- Wagner EH: Preventing decline in function. Evidence from randomized trials around the world. *West J Med* 1997, **167**:295-298.
- Pyne JM, Patterson TL, Kaplan RM, Gillin JC, Koch WL, Grant I: Assessment of the quality of life of patients with major depression. *Psychiatr Serv* 1997, **48**:224-230.
- Creed F, Morgan R, Fiddler M, Marshall S, Guthrie E, House A: Depression and anxiety impair health-related quality of life and are associated with increased costs in general medical inpatients. *Psychosomatics* 2002, **43**:302-309.
- Unutzer J, Katon W, Callahan CM, Williams JW Jr, Hunkeler E, Harpole L, Hoffing M, Della Penna RD, Noel PH, Lin EH, Areán PA, Hegel MT, Tang L, Belin TR, Oishi S, Langston C: Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. *JAMA* 2002, **288**:2836-2845.
- Callahan CM: Quality improvement research on late life depression in primary care. *Med Care* 2001, **39**:772-784.
- Bruce ML, Ten Have TR, Reynolds CF, Katz IR, Schulberg HC, Mulsant BH, Brown GK, McAvay GJ, Pearson JL, Alexopoulos GS: Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized controlled trial. *JAMA* 2004, **291**:1081-1091.
- O'Connor DW, Rosewarne R, Bruce A: Depression in primary care. 2: general practitioners' recognition of major depression in elderly patients. *Int Psychogeriatr* 2001, **13**:367-374.
- Cepoiu M, McCusker J, Cole MG, Sewitch M, Belzile E, Ciampi A: Recognition of depression by non-psychiatric physicians—a systematic literature review and meta-analysis. *J Gen Intern Med* 2008, **23**:25-36.
- Carson A, Margolin R: Depression in older patients with neurologic illness: causes, recognition, management. *Cleve Clin J Med* 2005, **72**(Suppl 3):S52-64.
- Lawhorne L: Depression in the older adult. *Prim Care* 2005, **32**:777-792.
- Spitzer RL, Kroenke K, Williams JB: Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA* 1999, **282**:1737-1744.
- Lyness JM, Noel TK, Cox C, King DA, Conwell Y, Caine ED: Screening for depression in elderly primary care patients. A comparison of the Center for Epidemiologic Studies-Depression Scale and the Geriatric Depression Scale. *Arch Intern Med* 1997, **157**:449-454.
- Kroenke K, Spitzer RL, Williams JB: The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001, **16**:606-613.
- Gilbody S, Richards D, Brealey S, Hewitt C: Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* 2007, **22**:1596-1602.
- Wittkamp KA, Naeije L, Schene AH, Huyser J, van Weert HC: Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen Hosp Psychiatry* 2007, **29**:388-395.
- Lowe B, Unutzer J, Callahan CM, Perkins AJ, Kroenke K: Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med Care* 2004, **42**:1194-1201.
- Centers for Disease Control and Prevention and National Association of Chronic Disease Directors: The State of Mental Health and Aging in America Issue Brief 1: What Do the Data Tell Us? 2008, Atlanta, GA.
- Lamers L, Jonkers CC, Bosma H, Penninx BW, Knottnerus JA, van Eijk JT: Summed score of the Patient Health Questionnaire-9 was a reliable and valid method for depression screening in chronically ill elderly patients. *J Clin Epidemiol* 2008, **61**:679-687.
- Lyness JM: Naturalistic outcomes of minor and subsyndromal depression in older primary care patients. *Int J Geriatr Psychiatry* 2008, **23**:773-781.
- Lyness JM, Chapman BP, McGriff J, Drayer R, Duberstein PR: One-year outcomes of minor and subsyndromal depression in older primary care patients. *Int Psychogeriatr* 2009, **21**:60-68.
- Ciechanowski P, Wagner E, Schmalting K, Schwartz S, Williams B, Diehr P, Kulzer J, Gray S, Collier C, LoGerfo J: Community-integrated home-based

- depression treatment in older adults: a randomized controlled trial. *JAMA* 2004, **291**:1569-1577.
27. Ackermann RT, Williams JW Jr: **Rational treatment choices for non-major depressions in primary care: an evidence-based review.** *J Gen Intern Med* 2002, **17**:293-301.
 28. Kroenke K: **Minor depression: midway between major depression and euthymia.** *Ann Intern Med* 2006, **144**:528-530.
 29. Spitzer RL, Williams JB, Gibbon M, First MB: **The Structured Clinical Interview for DSM-III-R (SCID). I: History, rationale, and description.** *Arch Gen Psychiatry* 1992, **49**:624-629.
 30. Lowe B, Kroenke K, Herzog W, Grafe K: **Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9).** *J Affect Disord* 2004, **81**:61-66.
 31. Kirshner B, Guyatt G: **A methodological framework for assessing health indices.** *J Chronic Dis* 1985, **38**:27-36.
 32. Lowe B, Kroenke K, Grafe K: **Detecting and monitoring depression with a two-item questionnaire (PHQ-2).** *J Psychosom Res* 2005, **58**:163-171.
 33. Kroenke K, Spitzer RL, Williams JB: **The Patient Health Questionnaire-2: validity of a two-item depression screener.** *Med Care* 2003, **41**:1284-1292.
 34. 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**:439-445.
 35. DeLong ER, DeLong DM, Clarke-Pearson DL: **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.** *Biometrics* 1988, **44**:837-845.
 36. Blank K, Gruman C, Robison JT: **Case-finding for depression in elderly people: balancing ease of administration with validity in varied treatment settings.** *J Gerontol A Biol Sci Med Sci* 2004, **59**:378-384.
 37. Wulsin L, Somoza E, Heck J: **The feasibility of using the Spanish PHQ-9 to screen for depression in primary care in Honduras.** *Prim Care Companion J Clin Psychiatry* 2002, **4**:191-195.
 38. Williams LS, Brizendine EJ, Plue L, Bakas T, Tu W, Hendrie H, Kroenke K: **Performance of the PHQ-9 as a screening tool for depression after stroke.** *Stroke* 2005, **36**:635-638.
 39. Fann JR, Bombardier CH, Dikmen S, Esselman P, Warms CA, Pelzer E, Rau H, Temkin N: **Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury.** *J Head Trauma Rehabil* 2005, **20**:501-511.
 40. Lotrakul M, Sumrithe S, Saipanish R: **Reliability and validity of the Thai version of the PHQ-9.** *BMC Psychiatry* 2008, **8**:46.
 41. Adewuya AO, Ola BA, Afolabi OO: **Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students.** *J Affect Disord* 2006, **96**:89-93.
 42. Gilbody S, Richards D, Barkham M: **Diagnosing depression in primary care using self-completed instruments: UK validation of PHQ-9 and CORE-OM.** *Br J Gen Pract* 2007, **57**:650-652.
 43. Hansson M, J Chotai J, Nordstrom A, Bodlund O: **Comparison of two self-rating scales to detect depression: HADS and PHQ-9.** *Br J Gen Pract* 2009, **59**:636-637.
 44. Wittkamp K, van Ravesteijn H, Baas K, van de Hoogen H, Schene A, Bindels P, Lucassen P, van de Lisdonk E, van Weert H: **The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care.** *Gen Hosp Psychiatry* 2009, **31**:451-459.
 45. Stafford L, Berk M, Jackson HJ: **Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients with coronary artery disease.** *Gen Hosp Psychiatry* 2007, **29**:417-424.
 46. Monahan PO, Shacham E, Reece M, Kroenke K, Ong'or WO, Omollo O, Yebei VN, Ojwang C: **Validity/reliability of PHQ-9 and PHQ-2 depression scales among adults living with HIV/AIDS in western Kenya.** *J Gen Intern Med* 2009, **24**:189-197.

Pre-publication history

The pre-publication history for this paper can be accessed here:
<http://www.biomedcentral.com/1471-2296/11/63/prepub>

doi:10.1186/1471-2296-11-63

Cite this article as: Phelan et al.: A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Family Practice* 2010 **11**:63.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

