Screening for Depression in Medical Settings with the Patient Health Questionnaire (PHQ): A Diagnostic Meta-Analysis

Simon Gilbody, MB DPhil MRCPsych, David Richards, PhD, Stephen Brealey, DPhil, and Catherine Hewitt, PhD

Department of Health Sciences, University of York, York YO10 5DD, UK.

OBJECTIVE: To summarize the psychometric properties of the PHQ2 and PHQ9 as screening instruments for depression.

INTERVENTIONS: We identified 17 validation studies conducted in primary care; medical outpatients; and specialist medical services (cardiology, gynecology, stroke, dermatology, head injury, and otolaryngology). Electronic databases from 1994 to February 2007 (MEDLINE, PsycLIT, EMBASE, CINAHL, Cochrane registers) plus study reference lists have been used for this study. Translations included US English, Dutch, Italian, Spanish, German and Arabic). Summary sensitivity, specificity, likelihood and diagnostic odds ratios (OR) against a gold standard (DSM-IV) Major Depressive Disorder (MDD) were calculated for each study. We used random effects bivariate meta-analysis at recommended cut points to produce summary receiver–operator characteristic (sROC) curves. We explored heterogeneity with metaregression.

MEASUREMENTS AND MAIN RESULTS: Fourteen studies (5,026 participants) validated the PHQ9 against MDD: sensitivity=0.80 (95% CI 0.71–0.87); specificity=0.92 (95% CI 0.88–0.95); positive likelihood ratio=10.12 (95% CI 6.52–15.67); negative likelihood ratio=0.22 (0.15 to 0.32). There was substantial heterogeneity (Diagnostic Odds Ratio heterogeneity I2=82%), which was not explained by study setting (primary care versus general hospital); method of scoring (cutoff $\geq$10 versus "diagnostic algorithm"); or study quality (blinded versus unblinded). The diagnostic validity of the PHQ2 was only validated in 3 studies and showed wide variability in sensitivity.

CONCLUSIONS: The PHQ9 is acceptable, and as good as longer clinician-administered instruments in a range of settings, countries, and populations. More research is needed to validate the PHQ2 to see if its diagnostic properties approach those of the PHQ9.

SG had the original idea for this meta-analysis, and produced the protocol, extracted data, undertook all analyses and produced initial and final drafts. DR, CH and SB execrated data and commented on all drafts of the paper.

Received April 23, 2007
Revised July 16, 2007
Accepted July 19, 2007
Published online September 14, 2007

KEY WORDS: depression; screening; questionnaire; psychometrics.
J Gen Intern Med 22(11):1596–602
DOI: 10.1007/s11606-007-0333-y

INTRODUCTION

Depression affects between 5% and 10% of individuals in primary care, but is only recognized in around 50% of cases.[1] Similar problems are also found in general hospital settings, where there is substantial under identification and unmet need for mental health services.[2] Depression is associated with personal suffering and decrements in quality of life and functioning.[3] Patients with unrecognized depression consult with their physician more frequently, and consume greater health care resources.[4] The presence of depression in conjunction with physical illness also adversely affects the outcome of both disorders.[2] Screening and case finding has been proposed to improve the recognition and management of depression.[5,6] To be acceptable in practice, instruments must be valid, reliable, brief, and easy to use.[7]

Previous research in this area includes a 2002 review by Williams and colleagues,[8] who found a number of instruments to be available for use in primary care settings, with acceptable psychometric properties (median sensitivity for major depression=85%, range 50–97%; median specificity=74%, range 51–98%). This is the benchmark against which all new instruments should be judged.

Two instruments that have been recently introduced are the two- and nine-item self-administered Patient Health Questionnaires—PHQ29 and PHQ9.[10,11] The specific items on these instruments are derived from the PRIME-MD interview schedules and are designed to establish DSM-IV criteria-based psychiatric diagnoses.[10] These instruments are brief, acceptable to patients, and can be self-administered. However, the diagnostic properties of both the PHQ2 and PHQ9 have yet to be summarized in any systematic manner.

In this article we apply systematic review and meta-analytic techniques to summarize the diagnostic properties of the Patient Health Questionnaires for depression.[12–14]

METHODS

Study Design and Setting. We included all cross-sectional validation studies[15] of the PHQ in primary care, community and hospital settings among adults.

Condition and Reference Test. We sought studies reporting the ability of the PHQ to detect depressive disorders. Disorders had to be defined according to standard classificatory systems, such as the International Classification of Diseases (ICD)[16] or Diagnostic and Statistical Manual of Mental Disorders (DSM).[17] Gold standard diagnoses had to be made using a standardized diagnostic interview schedule, such as the Structured Clinical Interview for DSM (SCID),[18] or the Diagnostic Interview Schedule.[19] Clinician-rated diagnoses without reference to specific structured diagnostic schedule were not included.

Diagnostic and Screening Instrument Cut Points. We sought studies examining the diagnostic properties of the self-administered two-item depression Patient Health Questionnaire (PHQ2)[9] and self-administered nine-item depression Patient Health Questionnaire (PHQ9).[10] The PHQ measures were developed within the PRIME-MD set of instruments and scales, and were designed for use in primary care and nonpsychiatric settings.[20] The nine-item PHQ contains items derived from the DSM-IV classification system pertain to: (1) anhedonia, (2) depressed mood, (3) trouble sleeping, (4) feeling tired, (5) change in appetite, (6) guilt or worthlessness, (7) trouble concentrating, (8) feeling slowed down or restless, (9) suicidal thoughts. The two-item PHQ includes only 2 questions pertaining to: (1) anhedonia, and (2) low mood.

We included English language versions and other languages. To maximize the available data and ensure consistency between studies, we used the most consistently reported and recommended cut point (10 or above for the PHQ9, and 3 or above for the PHQ2[9]). These cut points are the recommended scores to alert the physician to a significant depression requiring active intervention or surveillance (http://www.depression-primary-care.org/). Some studies also presented an alternative method scoring the PHQ9 using a "diagnostic algorithm"[10] (requiring 5 or more of the 9 depressive symptoms criteria to have been present "more than half the days" in the past 2 weeks, and 1 of the symptoms is depressed mood or anhedonia). The diagnostic algorithm also gives a score of $\geq$10 for the PHQ9. We also examined whether the use of this algorithm altered the diagnostic performance using metaregression and sensitivity analysis methods. Where alternative cut points were presented, we recorded these and reported whether alternative cut points performed better than the recommended cut point. We did not, however, pool alternate cut points in our overall meta-analysis, given the paucity of these data.

Search Strategy. We searched the following databases from 1994 (following publication of the initial validation studies) to February 2007: MEDLINE; EMBASE; PsycINFO; CINAHL. We constructed search terms from a series of depression-specific terms developed by the Cochrane Depression and Anxiety group,[21] and used a series of free-text terms to capture any publications that mentioned the PHQ and PRIME-MD instruments by name in their abstracts. We also searched specific internet sites related to the PHQ (http://www.depression-primarycare.org/) and scrutinized reference lists to include studies.

Data Extraction and Quality Assessment. Data were independently extracted by at least 2 researchers. We took care to avoid the "double counting" of evidence, particularly where the same first authors were quoted in several validation studies. We first constructed 2×2 tables for all studies at recommended cut off points. From these we calculated sensitivity, specificity, likelihood ratios (positive and negative).[22] The likelihood ratio represents a measure of the predictive ability of a test, which, unlike positive predictive value, is a fundamental predictive attribute of the instrument that does not vary according to the baseline prevalence of the disorder in question.[23] We also calculated the diagnostic odds ratio (DOR): the ratio of the odds of a positive test among those with the disorder to the odds of a positive result among those without the disorder.[13,24] This is the recommended metric in diagnostic meta-analyses.[12] We assessed study quality in line with accepted guidelines.[25] In particular, we sought information on the application of a diagnostic standard independent of the knowledge of scores on the PHQ instrument ("blinded"). Blinding is a potential source of bias within cross-sectional validation studies, as foreknowledge of test scores by those applying a diagnostic gold standard can create an exaggerated level of agreement.[25]

Data Synthesis and Meta-analysis. We undertook a bivariate diagnostic meta-analysis to obtain pooled estimates of: sensitivity and specificity; positive and negative likelihood ratios; and a summary diagnostic odds ratio. Briefly, this method fits a 2-level model, with independent binomial distributions for the true positives and true negatives conditional on the sensitivity and specificity in each study, and a bivariate normal model for the logit transforms of sensitivity and specificity between studies.[26] Our analysis used the generalized linear mixed model approach to bivariate meta-analysis.[27]

Receiver Operator Characteristic (ROC) curves are the most informative way of representing the inherent trade-offs between sensitivity and specificity for a test or diagnostic instrument.[28] We therefore created a single plot of sensitivity and specificity in ROC space, summarizing each study, weighted by study size. Summary Receiver Operator Characteristic curves (sROC)[29] were then constructed using the bivariate model to produce a 95% confidence ellipse within ROC space.[26] Unlike a traditional ROC plot that explores the effect of varying thresholds on sensitivity and specificity in a single study, each data point in the summary ROC space represents a separate study.

Between-study heterogeneity was assessed using the I^2 statistic of the pooled diagnostic odds ratio,[30] which describes the percentage of total variation across studies that is caused by heterogeneity rather than chance. The I^2 statistic has several advantages over other measures of heterogeneity (such as chi-square), including greater statistical power to detect clinical heterogeneity when fewer studies are available. As a guide, I^2 values of 25% may be considered "low", 50% "moderate" and 75%, "high". Where there was significant between-study heterogeneity, we sought to explore the causes of this heterogeneity. We first visually inspected our sROC plot to identify those studies that lay outside of the 95% confidence ellipse. We also undertook a metaregression analysis of a logit diagnostic odds ratio model using potential predictive covariates.[31] A priori causes of heterogeneity were:

1. Study quality—particularly the application of a diagnostic standard blind to PHQ score;
2. Study setting (primary care/community versus general hospital);

Table 1. Population and Design Features of Cross-Sectional Patient Health Questionnaire Validation Studies

Study	Setting, instrument and language	Diagnostic standard	Diagnosis blind to PHQ	Age, sex, sample size, and % depressed
Kroenke 2001[11]	PHQ9 US internal med, family practice, and Ob-Gyn clinics US English Diagnosis using algorithm	DSM SCID MDD All depressive disorders	Yes	Age=46 and 31 years Female= 66% & 100% *N*=580 MDD=41 (7.1%)
Lowe 2004[35,52]	PHQ9 German hospital outpatient and family practice clinics German translation Diagnosis using cut point ≥10[52] and algorithm35	DSM[35] and ICD[52] SCID MDD Any mood disorder	Yes	Age=41.7 years Female=67% *N*=501 MDD=66 (13.2%)
Fann 2005[36]	PHQ9 Head trauma US English Diagnosis using algorithm and cut point ≥10	DSM SCID MDD	No/ unclear	Age=42 years Female=30% *N*=135 MDD=22 (16.3%)
Watnick 2005[37]	PHQ9 Renal outpatients US English Diagnosis using cut point ≥10	DSM SCID MDD	Yes	Age=63 years Female=58% *N*=62 MDD=12 (23.1%)
Picardi 2005[38]	PHQ9 Dermatology outpatients Italian translation Diagnosis using algorithm	DSM SCID MDD All depressive disorders	Yes	Age=37.5 years Female=56% *N*=141 MDD=12 (8.5%)
Williams 2001 [39]	**PHQ9** US Stroke patients US English Diagnosis using cut point ≥10	DSM SCID MDD Any depressive disorder	No/ unclear	Age=unclear Female=unclear *N*=316 MDD=106 (33.5%)
Wulsin 2002 [40]	PHQ9 Honduran rural primary care Honduran-Spanish translation Diagnosis using cut point ≥10	DSM SCID MDD	No/ unclear	Age=32.2 years Female=100% N=34 MDD=13 (38.2%)
Persoons 2003[41]	PHQ9 Otolaryngology outpatients Dutch version Diagnosis using algorithm	DSM MINI interview MDD Any mood disorder	Yes	Age=48.2 years Female=65.6% *N*=97 MDD=16 (16.5%)
Becker 2002[42]	PHQ9 Saudi Primary care Arabic version Diagnosis using algorithm	DSM SCID Any mood disorder	No/ unclear	Age=75% under 50 years Female=54.1% *N*=431 ADD= 86 (20.0%)
Diez-Quevedo 2001[43]	PHQ9 Spanish general hospital patients Spanish version Diagnosis using algorithm	DSM SCID MDD Any mood disorder	Yes	Age=43.0 years Female=45.6% *N*=1003 MDD=148 (14.8%)
McManus 2005[45]	PHQ2 & PHQ9 US cardiology outpatients. US English Diagnosis using cut point ≥3 (PHQ2) and ≥10 (PHQ9)	DSM DIS MDD	No/ unclear	Age=67 years Female=18% *N*=1024 MDD=224 (21.9%)
Kroenke 2003[9]	PHQ2 US primary care and gynecology patients US English Diagnosis using cut point ≥3	DSM SCID MDD Any mood disorder	Yes	Age=46 & 31 years Female= 66% and 100% *N*=580 MDD= 7.1%
Henkel 2004[44]	PHQ9 German primary care clinics German translation Diagnosis using algorithm	DSM CIDI MDD Any mood disorder	Yes	Age=46 & 31 years Female= 74% *N*=431MDD=10.0%
Loewe 2005[46]	PHQ2 German hospital outpatient and family practice clinics German translation Diagnosis using cut point ≥3	DSM SCID MDD Any mood disorder	Yes	Age=41.7 years Female=67% *N*=501 MDD=13.2%
Eack 2006[47]	PHQ9 Mothers of children attending a US community child psychiatric unit US English Diagnosis using algorithm	DSM SCID MDD ADD	No/ unclear	Age=39.2 years Female=100% *N*=50 MDD=28%
Adewuya 2006[48]	PHQ9 Nigerian University students (non-clinical)English Diagnosis using cut point ≥10	DSM MINI MDD	No/ unclear	Age=24.8 years Female=42% *N*=512 MDD=2.5%
Gilbody 2007[49]	PHQ9 UK Primary care English Diagnosis using cut point ≥10	DSM SCID MDD	Yes	Age=42 years Female=71% *N*=96 MDD=38%

Abbreviations: DSM = Diagnostic and Statistic Manual; SCID = Structured Clinical Interview for DSM-III-R; MDD = major depressive disorder; DIS = Diagnostic Interview Schedule; ICD = International Classification of Diseases; CIDI = Composite International Diagnostic Interview; ADD = Any Depressive Disorder; MINI = Mini-International Neuropsychiatric Interview; PHQ = Patient Health Questionnaire

3. Baseline prevalence of depression in the screened population, as a proxy measure of the spectrum of severity of disorder within the screened population.
4. Method of scoring (cut using raw score ≥10 versus diagnostic algorithm)

If these items were important sources of heterogeneity, then they would be predictive in a metaregression analysis, and would reduce the level of between-study heterogeneity in metaregression model. For dichotomous predictor variables, this metaregression model produced a "ratio of diagnostic odds ratios", where deviation from 1 suggested difference in pooled estimates according to a predictor variable. All metaregression analyses were conducted using a permutation test to minimize type 1 errors, with 1,000 Monte Carlo replications to generate *p* values.[32]

Finally, publication and small study bias was examined using Begg funnel plots[33] and by testing for funnel plot asymmetry using Egger's weighted regression test.[34] Analyses were conducted using Stata version 9, with the metandi, metabias, and metareg user-written commands.

RESULTS

Our searches identified 749 potential studies. From these, we found 17 validation studies meeting our inclusion criteria and

Table 2. Diagnostic Attributes for PHQ9 and PHQ2 Instruments Against a Standardized Diagnosis of DSM Major Depressive Disorder

Population	Sensitivity	Specificity	LR+	LR-ve	DOR
PHQ9 All studies (n=14 studies)[11,35–41,43–45,47–49]	0.80 (0.71–0.87)	0.92 (0.88–0.95)	10.12 (6.52–15.7)	0.22 (0.15–0.32)	46.12 (23.7–89.6)
Primary care and community[11,35,40,44,47–49] (n=7 studies)	0.81 (0.72–0.88)	0.92 (0.83–0.97)	11.30 (4.56–28.0)	0.20 (0.13–0.31)	56.44 (18.9–169)
All hospital specialities (n=7 studies)	0.78 (0.64–0.88)	0.91 (0.90–0.92)	8.72 (6.94–11.0)	0.24 (0.14–0.41)	36.38 (17.5–75.8)
Cardiology[45]	0.54 (0.47–0.61)	0.90 (0.88–0.92)	5.40 (4.25–6.87)	0.51 (0.44–0.59)	10.57 (7.45–15.00)
Renal dialysis[37]	0.92 (0.61–0.99)	0.92 (0.80–0.98)	12.22 (4.06–36.78)	0.09 (0.01–0.59)	135.67 (12.8–1438)
Otolaryngology[41];	0.69 (0.41–0.89)	0.95 (0.88–0.99)	13.92 (5.07–38.26)	0.33 (0.16–0.68)	42.35 (9.8–182)
brain injury[36];	0.86 (0.65–0.97)	0.90 (0.83–0.95)	8.87 (4.94–15.93)	0.15 (0.05–0.43)	58.73 (15.0–230)
Stroke services[39];	0.91 (0.83–0.95)	0.89 (0.84–0.93)	7.93 (5.41–11.61)	0.11 (0.06–0.19)	74.40 (34.2–162)
Dermatology;[38]	0.50 (0.21–0.79)	0.91 (0.84–0.95)	5.38 (2.46–11.74)	0.55 (0.31–0.97)	9.75 (2.72–35.00)
General medical outpatients[43]	0.84 (0.77–0.89)	0.92 (0.90–0.94)	10.54 (8.30–13.38)	0.18 (0.12–0.25)	59.8 (36.2–98.8)
PHQ2 All studies (n=3 studies)[9,45,46]					
Primary care studies[9] (n=1)	0.83 (0.68–0.93)	0.83 (0.68–0.93)	8.28 (6.20–11.05)	0.19 (0.10–0.37)	43.62 (18.45–103.16)
General hospital studies (n=2)	0.51 (0.45–0.56)	0.87 (0.85–0.89)	4.29 (3.46–5.31)	0.34 (0.07–1.71)	12.79 (3.87–42.24)
Cardiology[45]	0.39 (0.32–0.46)	0.92 (0.90–0.94)	4.86 (3.65–6.47)	0.67 (0.60–0.74)	7.30 (5.04–10.58)
General medical outpatients [46,47]	0.87 (0.77–0.94)	0.78 (0.74–0.82)	3.96 (3.26–4.81)	0.16 (0.09–0.30)	24.36 (11.69–50.73)

Abbreviations: PHQ =Patient Health Questionnaire; DSM = Diagnostic and Statistic Manual; LR+ = Likelihood ratio for a positive test; LR-ve = Likelihood ratio for a negative test; DOR=Diagnostic odds ratio

with sufficient data (reported in 18 publications[9,11,35–49]—see Table 1.) Fourteen studies examined the diagnostic attributes of the PHQ9,[11,35–45,47–49] and 3 studies examined the diagnostic attributes of the PHQ2[9,45,46] (three studies reported the diagnostic properties of both the PHQ2 and PHQ9—one in a single report[45] and 2 each in 2 separate reports.[9,11,35,46]

Eight studies were conducted in primary care or community settings,[9,11,40,42,44,47–49] 3 in general medical outpatients,[35,43,46] and 6 in specialist hospital services (dermatology[38]; stroke services[39]; brain injury[36]; otolaryngology[41]; renal dialysis services[37]; cardiology[45]). Mean age ranged from 25 to 67 years. Within these studies, there was a large variation in the proportion of participants in the screened population with depressive disorder, as ascertained by gold standard diagnosis. Prevalence ranged from 2.5%[48] to 38%.[40] Studies used US English[9,11,36,37,39,45,47–49]; German[35,44,46]; Italian[38]; Spanish[40,43]; Dutch[41]; and Arabic[42] versions of the PHQ9—see Table 1 for details.

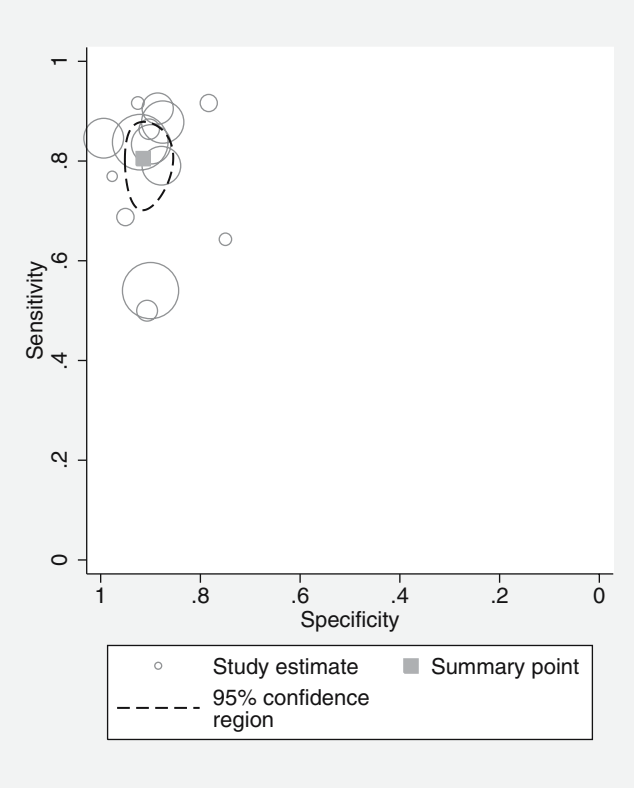

Figure 1 PHQ9 summary ROC plot of diagnosis of major depressive disorder at cutoff ≥10 or by "diagnostic algorithm." Pooled co-distribution of sensitivity and specificity using a bivariate meta-analysis. Individual point estimates represent single studies, with size of circle proportionate to study sample size.

All studies made a reference to a DSM diagnosis of depression, established according to standardized methods by a trained research worker or mental health professional. Interview schedules included the SCID,[18] CIDI,[50] MINI,[51] or the Diagnostic Interview Schedule.[19] However, there was no specific mention of blinding of gold standard assessors to the results of the PHQ instruments in 7 studies.[36,39,40,42,45,47,48] All 17 studies reported the presence or absence of DSM Major Depressive Disorder (MDD). One study reported 2 separate gold standard validations on the same population[35,52]: 1 using DSM-IV[35] and 1 using ICD.[52] For the purposes of our meta-analysis, we use only the DSM major depressive disorder (MDD) diagnoses using recommended cut points/diagnostic algorithms to maintain diagnostic consistency between studies and to avoid "double counting" of evidence.

Meta-analysis of PHQ9 in Detecting Major Depressive Disorder

We pooled 14 studies (5,026 patients: 770 confirmed cases of major depressive disorder by DSM gold standard).[11,35–41,43–45,47–49] When we combined psychometric attributes across studies, we found a high level of between-study heterogeneity (combined diagnostic odds ratio I^2=82%) and no evidence of publication or small-study bias (p=0.21). Pooled sensitivity was 0.80 (95% CI 0.71–0.87) and specificity was 0.92 (95% CI 0.88–0.95)—Table 2. Likelihood ratio for a positive test was 10.12 (6.52 to 15.67), and the likelihood ratio for a negative test was 0.22 (0.15 to 0.32). When we summarized individual studies within ROC space, we found the majority of studies were gathered within an informative top left corner (Fig. 1). However, 3 studies were obvious "outliers": a study

Table 3. Metaregression Analysis of Sources of Potential Diagnostic Heterogeneity on the Performance of the PHQ9

Predictor variable	DOR1	DOR2	Beta-coefficient (95% CI)	*P* value	I^2 (%)
Study quality	$DOR_{blinded}$=38.66	$DOR_{unblinded}$=42.18	Ratio of DOR=0.96 (0.24–3.91)	.96	95
Study setting	$DOR_{primary\ care}$=44.19	$DOR_{gen\ hosp}$=35.98	Ratio of DOR=0.83 (0.19–3.65)	.79	98
Prevalence of depression	Na	Na	0.06 (0.01–41.03)	.37	98
Method of scoring	$DOR_{\geq 10}$=62.93	$DOR_{diag\ algorithm}$=26.14	Ratio of DOR=0.43 (0.12–1.54)	.17	96

Abbreviations: PHQ = Patient Health Questionnaire; DOR = Diagnostic odds ratio

of diagnostic properties of the PHQ9 in cardiology patients,[45] in dermatology patients,[38] and in mothers of children with behavioral disorders.[47] Each of these had relatively low sensitivity. However, the overall pooled result was not substantially altered by the inclusion and exclusion of these studies. When these outliers were omitted, the pooled sensitivity was 0.82 (95% CI 0.72–0.90). The omission of these studies substantially reduced the level of between-study heterogeneity from high (82%) to low (36%). Of the 2 hospital-based studies,[45,38] these results were not typical of other hospital-based studies and no clear difference in terms of patient populations was evident between these studies and other hospital-based studies. The community-based study[47] was slightly atypical in that it was a nonclinical population of mothers who were not themselves being referred for medical services.

Given the presence of between-study heterogeneity, we also conducted a metaregression. None of our a priori sources of heterogeneity were predictive when entered as covariates in our metaregression model. We found the diagnostic odds ratio similar in hospital settings compared to primary care settings (primary care vs hospital settings; beta-coefficient {ratio of diagnostic odds ratios}=0.83, 95% CI 0.19–3.64); p=0.79; I^2=98%). Primary care and hospital-based studies were also equally heterogenous (primary care I^2=90%; hospital-based I^2=88%). Diagnostic performance did not significantly vary according to method of scoring (cut point ≥10 versus "diagnostic algorithm" {ratio of diagnostic odds ratios}: beta=0.43, 95% CI 0.12–1.54; p=0.20). Similarly, the performance of the instrument did not significantly vary according to the baseline prevalence of depression (beta=0.06, 95% CI 0.01–41.03; p=0.37). For a detailed summary of diagnostic properties stratified by setting and patient population, see Tables 2 and 3.

Alternate Cut Points for the PHQ9

Six studies reported a range of cut points for the PHQ9 (Table 4).[11,35,36,48,49,52] The recommended cut point of ≥10 represented the optimum cut point in only one of these studies.[11] In 1 community (nonclinical) study,[48] the cut point of 9 represented the optimum (sensitivity=0.92, specificity=0.98). In a hospital-based study among patients with brain injury,[36] authors reported that ≥12 represented the optimum point. In a German community-based study, the cut point ≥11 represented the optimum and in a UK primary care study, ≥11 and ≥12 presented the optimum combinations of sensitivity/specificity.[49]

Properties of PHQ2 in Detecting Major Depressive Disorder

We identified 3 studies (2,124 patients: 336 confirmed cases of major depression).[9,45,46] We found a high level of between-study heterogeneity (DOR I^2=89%) and insufficient studies to test for publication bias. We, therefore, did not proceed to a full bivariate meta-analysis for the PHQ2, and instead present the results of individual studies separately.

Studies were conducted in primary care,[9] cardiology,[45] and general medical outpatients.[46] Sensitivity was good in 2 studies (0.83[9] and 0.87[46]), but was poor in the cardiology population (0.39[45]). It is worthy of note that this was the same study where the PHQ9 instrument had performed poorly. Despite this, the specificity was good (range 0.78–0.92).

DISCUSSION

This systematic overview of the diagnostic properties of the PHQ2 and PHQ9 instruments is the first to summarize recent validation work beyond the original population studies.[10,11] Our methods use recently developed multilevel meta-analytic methods, which have not (to our knowledge) hitherto been applied in psychometric studies. Previous overviews in this area have either not used quantitative synthesis methods,[8] or have applied methods that ignore the codependence of sensitivity and specificity.

We found 17 validation studies, with more than 5,000 participants where a diagnostic gold standard had been independently applied. Our main finding is that, for major depressive disorder, the PHQ9 has good[53] diagnostic properties, and was able to correctly diagnose major depression (sensitivity 92%) while being able to exclude this condition with some certainty (specificity 80%). Despite the clinical heterogeneity of studies in terms of settings (community, primary care, and a range of hospital specialities), the properties of the PHQ for major depression were relatively consistent between a range of settings and specialities. There were 3 obvious exceptions to this case, and despite our best efforts, we could find no consistent unifying feature to identify causes of this statistical heterogeneity. A further finding is that the PHQ9 performs well in a range of cultures and with a range of translations. These properties compare very well, or exceed the properties of clinician-administered instruments for major depression.[8]

The diagnostic properties of the PHQ9 were relatively good using either the ≥10 or "diagnostic algorithm" method, and our meta-analysis was unable to find a clear difference in performance according to which method was used. However, our meta-regression analysis may have lacked the statistical power to detect any systematic differences in test performance. The optimum diagnostic cut point was also called into question by the publication of sensitivity/specificity values in a subset of studies where the optimum was not always ≥10. It was clear that in some community-based studies, the cut point may be increased to ≥11 or ≥12 to obtain optimum specificity. However, we are cautious of this finding, as it may be that only studies where the optimum cut point varied from ≥10 chose to report

Table 4. Diagnostic Performance at Varying Cut Points for the PHQ9

Study	PHQ >=9	PHQ >=10	PHQ >=11	PHQ >=12	PHQ >=13
Kroenke 2001[11]	Sens.=0.95	Sens.=0.88	Sens.=0.83	Sens.=0.83	Sens.=0.78
	Spec.=0.84	Spec.=0.88	Spec.=0.89	Spec.=0.92	Spec.=0.93
Lowe 2004[35]	Not reported	Not reported	Sens.=0.98	Sens.=0.95	Sens.=0.88
			Spec.=0.80	Spec.=0.84	Spec.=0.87
Lowe 2004[52]ICD	Not reported	Sens.=0.90	Sens.=0.89	Sens.=0.86	Sens.=0.78
		Spec.=0.77	Spec.=0.80	Spec.=0.84	Spec.=0.88
Fann 2005[36]	Not reported	Sens=0.88	Not reported	Sens=0.85	Not reported
		Spec=0.92		Spec=0.94	
Adewuya 2006[48]	Sens.=0.92	Sens.=0.85	Sens.=0.62	Sens.=0.46	Not reported
	Spec.=0.98	Spec.=0.99	Spec.=0.99	Spec.=1.00	
Gilbody 2007[49]	Sens.=0.94	Sens.=0.92	Sens.=0.92	Sens.=0.92	Sens.=0.90
	Spec.=0.73	Spec.=0.78	Spec.=0.82	Spec.=0.85	Spec.=0.87

Abbreviations: PHQ = Patient Health Questionnaire; Sens = sensitivity; Spec = specificity

diagnostic properties at different points. The cut point of ≥10 still represents a diagnostic performance that exceeds most other published instruments.[54] Clinicians and researchers may therefore vary in their choice of cut point according to the clinical population, and the data in the present meta-analysis provide some empirical justification for this.

The methodological quality of several studies was poor according to our chosen criterion of blind application of a diagnostic gold standard. This raises a more general problem of the quality and reporting of diagnostic studies outlined in a recent consensus document.[55] We would hope that the introduction of uniform journal reporting of diagnostic studies (the STARD guidelines[55]) will improve both the conduct and reporting of studies in future diagnostic reviews of this and other studies.

We found substantial between-study variation, but were unable to fully explain this using the a priori sources of heterogeneity that we specified in our metaregression model. Our ability to explore this further was limited by the relatively small numbers of primary studies that we found. Other potential sources of heterogeneity might include: country of origin, language of translation, and type of diagnostic gold standard interview used. A larger number of validation studies will be required to allow further analyses to be performed with any statistical power, or without increasing the risk of finding "false positive" associations.

When we examined the psychometric properties of the PHQ2, there were far fewer studies with a diagnostic gold standard. The psychometric attributes of this instrument need to be subject to a much wider range of validation studies across different clinical settings and patient populations before the validity of this brief measure can be assumed. The PHQ2 seems a promising instrument, and if its validation approaches that of the PHQ9, its brevity will make it a very attractive instrument to use in routine settings. More research in this area is required as a matter of urgency.

There is substantial under-recognition of depressive disorders in primary care and hospital settings, and the PHQ has now emerged a candidate instrument, which can be adopted in routine practice, within a variety of settings. It compares well with longer or clinician-administered instruments. However, caution should be exercised in implementing screening procedures, in the hope that instruments alone will improve the quality of care and patient outcomes.[56] Randomized studies of the introduction of depression screening instruments into clinical practice generally fail to demonstrate any substantial benefit in terms of improved patient management and outcomes. One of the main reasons for the negative impact of screening instruments is likely to be their low predictive value, given the relatively low prevalence of depression in primary care populations.[56] Even highly sensitive and specific screening tests are likely to be wrong more often than they are right in nonspecialist settings with a low base rate of depression. Recent recommendations from the US Preventive Services Task Force highlight that screening tools should only be implemented alongside enhancements of patient care[5] to ensure that commensurate high-quality care is available for those with hitherto unrecognized depression. Candidate depression enhancements include collaborative care, and successful programs often include case finding instruments such as the PHQ.[57]

Acknowledgments: *We are grateful to Dr Peter Bower for comments on an earlier draft of the manuscript. We also thank authors for providing unpublished data, and answering queries about study design. There is no external or internal funding for this project.*

Conflict of interest: *None disclosed.*

Corresponding Author: *Professor Simon Gilbody, MB DPhil MRCPsych;Department of Health Sciences, University of York, York YO10 5DD, UK (e-mail: sg519@york.ac.uk).*

REFERENCES

1. **Simon G, Von Korff M.** Recognition and management of depression in primary care. Arch Fam Med. 1995;4:99–105.
2. **Katon W, Ciechanowski P.** Impact of major depression on chronic medical illness. J Psychosom Res. 2002;53:859–63.
3. **Wells KB, Stewart A, Hays RD, et al.** The functioning and well-being of depressed patients. Results from the Medical Outcomes Study. JAMA. 1989;262(7):914–9.
4. **Simon GE, Chisholm D, Treglia M, Bushnell D.** Course of depression, health services costs, and work productivity in an international primary care study. Gen Hosp Psych. 2002;24(5):328–35.
5. **Pignone MP, Gaynes BN, Rushton JL, et al.** Screening for depression in adults: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med. 2002;136:765–76.
6. Agency for Healthcare Research and Quality. Screening for Depression: Systematic Evidence Review Number 6. Rockville MD: AHRQ, 2002.
7. **Street RL, Jr., Gold WR, McDowell T.** Using health status surveys in medical consultations. Med Care. 1994;32(7):732–44.
8. **Williams JW, Pignone M, Ramirez G, Stellato CP.** Identifying depression in primary care: a literature synthesis of case-finding instruments. Gen Hosp Psych. 2002;24:225–37.

9. **Kroenke K, Spitzer RL, Williams JB.** The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care. 2003;41:1284–92.
10. **Spitzer RL, Kroenke K, Williams JBW.** 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;282:1737–44.
11. **Kroenke K, Spitzer RL, Williams JB.** The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16:606–13.
12. **Deeks J.** Evaluations of diagnostic and screening tests. In: Egger M, Davey Smith G, Altman DG, eds. Systematic Reviews in Health Care. London: BMJ Books, 2000:248–82.
13. **Deville WL, Buntinx F, Bouter LM, et al.** Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol. 2002;2:9.
14. **Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J.** Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess. 2004;8:1–234.
15. **Knottnerus JA, Muris JW.** Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol. 2003;56:1118–28.
16. World Health Organisation. International Statistical Classification of Diseases and Related Health Problems—10th Revision. Geneva: WHO, 1990.
17. American Psychiatric Association. Diagnostic and Statistical Manual—4th Edition. Washington DC: American Psychiatric Association, 1994.
18. **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(8):624–9.
19. **Robins LN, Helzer JE, Croughan J, Ratcliff KS.** National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. Arch Gen Psychiatry. 1981;38:381–9.
20. **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;272:1749–56.
21. **Churchill R, Hunot V, McGuire H.** Cochrane Depression Anxiety and Neurosis Group. Cochrane Library 2004;2.
22. **Sackett DL, Haynes RB, Guyatt GH, Tugwell P.** Clinical Epidemiology: A basic science for clinical medicine. Boston, MA.: Little, Brown and Company, 1991.
23. **Sackett DL, Haynes RB.** Evidence base of clinical diagnosis: the architecture of diagnostic research. BMJ. 2002;324:539–41.
24. **Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM.** The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003;56:1129–35.
25. **Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J.** Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140(3):189–202.
26. **Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH.** Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58:982–90.
27. **Chu H, Cole SR.** Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol. 2006;59:1331–32.
28. **Knottnerus JA, ed.** The evidence base of clinical diagnosis. London: BMJ Publishing, 2002.
29. **Walter SD.** Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med. 2002;21:1237–56.
30. **Higgins JP, Thompson SG, Deeks JJ, Altman DG.** Measuring inconsistency in meta-analyses. BMJ. 2003;327:557–60.
31. **Thompson SG, Higgins JP.** How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002;21:1559–73.
32. **Higgins JPT, Thompson SG.** Controlling the risk of spurious findings from meta-regression. Stat Med. 2004;23:1663–82.
33. **Begg CB.** Publication bias. In: Cooper H, Hedges LV, eds. The handbook of research synthesis. New York: Russell Sage Foundation, 1994: 399–409.
34. **Egger M, Davey-Smith G, Schneider M, Minder C.** Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315:629–34.
35. **Lowe B, Spitzer RL, Grafe K, et al.** Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. J Affect Disord. 2004;78:131–40.
36. **Fann JR, Bombardier CH, Dikmen S, et al.** Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. J Head Trauma Rehabil. 2005;20:501–11.
37. **Watnick S, Wang PL, Demadura T, Ganzini L.** Validation of 2 depression screening tools in dialysis patients. Am J Kidney Dis. 2005;46:919–24.
38. **Picardi A, Adler DA, Abeni D, et al.** Screening for depressive disorders in patients with skin diseases: a comparison of three screeners. Acta Derm Venereol. 2005;85:414–9.
39. **Williams LS, Brizendine EJ, Plue L, et al.** Performance of the PHQ-9 as a screening tool for depression after stroke. Stroke. 2005;36:635–8.
40. **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 Psychiatr. 2002;4:191–5.
41. **Persoons P, Luyckx K, Desloovere C, Vandenberghe J, Fischler B.** Anxiety and mood disorders in otorhinolaryngology outpatients presenting with dizziness: validation of the self-administered PRIME-MD Patient Health Questionnaire and epidemiology. Gen Hosp Psych. 2003;25:316–23.
42. **Becker S, Al Zaid K, Al Faris E.** Screening for somatization and depression in Saudi Arabia: a validation study of the PHQ in primary care. Int J Psychiatry Med. 2002;32:271–83.
43. **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;63:679–86.
44. **Henkel V, Mergl R, Kohnen R, Allgaier A, Möller H, Hegerl U.** Use of brief depression screening tools in primary care: consideration of heterogeneity in performance in different patient groups. Gen Hosp Psych. 2004;26(3):190–8.
45. **McManus D, Pipkin SS, Whooley MA.** Screening for depression in patients with coronary heart disease (data from the Heart and Soul Study). Am J Cardiol. 2005;96:1076–81.
46. **Lowe B, Kroenke K, Grafe K.** Detecting and monitoring depression with a two-item questionnaire (PHQ-2). J Psychosom Res. 2005;58:163–71.
47. **Eack S, Greeno CG, Lee BJ.** Limitations of the Patient Health Questionnaire in identifying anxiety and depression in community mental health: many cases are undetected. Res Soc Work Pract. 2006;16:625–31.
48. **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.
49. **Gilbody S, Richards D, Barkham M.** Diagnosing depression in primary care using self-completed instruments: a UK validation of the PHQ9 and CORE-OM. Br J Gen Pract. 2007;57(541):65–652.
50. **Andrews G, Peters L.** The psychometric properties of the Composite International Diagnostic Interview. Soc Psychiatry Psychiatr Epidemiol. 1998;33:80–8.
51. **Sheehan DV, Lecrubier Y, Sheehan KH, et al.** The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry. 1998;59(Suppl 20):22–33.
52. **Lowe B, Grafe 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;73:386–90.
53. **Streiner D, Norman G.** Health Measurement Scales: A practical guide to their development and use, 3rd ed. Oxford, UK.: Oxford University Press, 2003.
54. **Williams JW, Noel PH, Cordes JA, Ramirez G, Pignone M.** Is this patient clinically depressed? JAMA. 2002;287:1160–70.
55. **Bossuyt PM, Reitsma JB, Bruns DE.** Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem. 2003;49:1–6.
56. **Gilbody S, Sheldon T, Wessely S.** Should we screen for depression? BMJ. 2006;332(7548):1027–30.
57. **Unutzer J, Katon W, Callahan CM, et al.** Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. JAMA. 2003;288:2836–45.