BJPsych Open (2016)
2, 127–138. doi: 10.1192/bjpo.bp.115.001685

Review

Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies

Alex J. Mitchell, Motahare Yadegarfar, John Gill and Brendon Stubbs

Background
The Patient Health Questionnaire (PHQ) is the most commonly used measure to screen for depression in primary care but there is still lack of clarity about its accuracy and optimal scoring method.

Aims
To determine via meta-analysis the diagnostic accuracy of the PHQ-9-linear, PHQ-9-algorithm and PHQ-2 questions to detect major depressive disorder (MDD) among adults.

Method
We systematically searched major electronic databases from inception until June 2015. Articles were included that reported the accuracy of PHQ-9 or PHQ-2 questions for diagnosing MDD in primary care defined according to standard classification systems. We carried out a meta-analysis, meta-regression, moderator and sensitivity analysis.

Results
Overall, 26 publications reporting on 40 individual studies were included representing 26 902 people (median 502, s.d.=693.7) including 14 760 unique adults of whom 14.3% had MDD. The methodological quality of the included articles was acceptable. The meta-analytic area under the receiver operating characteristic curve of the PHQ-9-linear and the PHQ-2 was significantly higher than the PHQ-9-algorithm, a difference that was maintained in head-to-head meta-analysis of studies. Our best estimates of sensitivity and specificity were 81.3% (95% CI 71.6–89.3) and 85.3% (95% CI 81.0–89.1), 56.8% (95% CI 41.2–71.8) and 93.3% (95% CI 87.5–97.3) and 89.3% (95% CI 81.5–95.1) and 75.9% (95% CI 70.1–81.3) for the PHQ-9-linear, PHQ-9-algorithm and PHQ-2 respectively. For case finding (ruling in a diagnosis), none of the methods were suitable but for screening (ruling out non-cases), all methods were encouraging with good clinical utility, although the cut-off threshold must be carefully chosen.

Conclusions
The PHQ can be used as an initial first step assessment in primary care and the PHQ-2 is adequate for this purpose with good acceptability. However, neither the PHQ-2 nor the PHQ-9 can be used to confirm a clinical diagnosis (case finding).

Declaration of interest
None.

Major depressive disorder (MDD) is a serious, disabling condition that is often comorbid with other medical presentations.[1–4] Most care for depression is delivered by general practitioners (GPs) and individually many GPs have considerable experience in managing depression.[5] Approximately 7% of all consultations in primary care are for depression.[6] Yet, clinicians find it challenging to precisely diagnose depression and often overestimate or underestimate levels of distress of their patients sometimes resulting in false-positive or false-negative diagnoses.[7] Indeed, GPs are typically able to detect about half of true cases of depression on a one-off visit[1] and once diagnosed not all patients with depression receive adequate timely care.[8] Although under-detection can lead to inadequate treatment,[9] over-detection (misidentification) can lead to inappropriate treatment.[9,10] For example, in the Baltimore Epidemiologic Catchment Area Study, 38% of antidepressant users never met the criteria for MDD, obsessive–compulsive disorder, panic disorder, social phobia or generalised anxiety disorder in their lifetime.[10] Mitchell *et al*[1] suggested that this could become a particular problem in routine care where prevalence rates are modest when false positives can outnumber false negatives.

Given that many clinicians have highlighted the difficulties in the timely diagnosis of depression[11] and that depression care is often inadequate,[12–14] the use of screening tools in routine care has been suggested by some as possibly beneficial by enhancing diagnosis-as-usual. Screening is most usefully defined as the systematic application of a test to rule out those without a condition and case finding most usefully defined as the systematic application of a test to confirm those with a condition.[15] Screening and case finding have been proposed as solutions adopted into the UK primary care quality outcomes framework (QoF).[16] The use of short screening questionnaires (<5 min) and ultra-short questionnaires (<2 min) may improve the recognition of depression if such tests are accurate, acceptable and implemented.[17,18] Of all the possible tools for depression, the depression module of the Patient Health Questionnaire (PHQ-9) is the most popular current tool which has three main formats.

1 The PHQ-9 (PHQ-9-linear) scored by simple addition and at a threshold of 10 or higher had a sensitivity of 88% and a specificity of 88% for detecting MDD in the initial validation study.[19]
2 The PHQ-9 (PHQ-9-algorithm) scored by the algorithm suggested in DSM-IV for MDD (the DSM algorithm method requires at least five symptoms rated as at least 2 (more than half the days) (>0 for the suicidal ideation item) plus at least one of the symptoms scored as at least 2 is either loss of interest or pleasure or depressed mood all present for 2 weeks

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

or more and associated with distress or dysfunction). As this follows the rules of DSM-IV more precisely, it is anticipated that this method should be the most accurate.

3 The PHQ-2 is the 2-item version utilising only the first two questions, namely loss of interest and low mood for the past 2 weeks, scored by simple linear scoring using a threshold of 2 or higher.[20]

An adaptation of the PHQ-2 also exists where the main modification is the duration of questioning which is over the past month rather than 2 weeks. This is known as the Whooley questions after the original author.

Yet, it is important to acknowledge that the value of screening and severity assessment has been disputed both in the literature and in clinical practice. Some authors have stated that routine use of depression tools should identify patients with either previously unrecognised MDD or untreated MDD (in effect a demonstration of added value)[21] but policy recommendations and guidelines have been inconsistent. In 2009, the United States Preventive Services Task Force (USPSTF) recommended routine depression screening in primary care settings with follow-up.[22] This recommendation has recently been revised and extended.[23] In the UK, the national guidelines have reversed their advice[24] and the most recent draft guidance state there is little convincing evidence that depression screening will reduce the number of patients with depression or improve depression symptoms.[25] GPs in the UK have been less enthusiastic than patients about routine use of depression scales,[26] leading to the removal of depression screening incentives from the UK QoF. In 2013, the Canadian CTFPHC reconsidered its earlier guideline and also recommended against screening adults for depression in primary care settings.[27] Thus, although some still advocate screening for depression, others do not and the argument has become polarised.[23] Few are putting forth the argument that screening might work in some circumstances or that further evidence is required from high-quality studies, leading to observers to suggest that this is a form of confirmation bias from either side defending an entrenched position.[28]

Four previous meta-analyses have been conducted on the accuracy of the PHQ-9 but none have specifically been conducted in primary care.[29–32] One previous meta-analysis has been conducted on the PHQ-2/Whooley questions but is considerably out of date.[17] Thus, the primary objective is to conduct a meta-analysis to determine the diagnostic accuracy of the PHQ-9-linear, PHQ-9-algorithm and PHQ-2 questions to detect MDD among adults.

Method

This systematic review was conducted following a predetermined but unpublished protocol.

Inclusion and exclusion criteria

We included studies that reported the accuracy of PHQ-9/PHQ-2 questions for diagnosing MDD in primary care. The setting had to be mostly primary care (but not exclusively, containing >50% of primary care patients) and we identified one study in two publications with mixed recruitment.[33,34] Studies focusing on one single medical condition in primary care were excluded.[35] The studies had to provide sufficient data to allow us to calculate contingency tables or had to be supplied by authors. We only included studies that defined MDD according to standard classification systems such as the ICD or the DSM using a standardised diagnostic interview schedule (Mini International Neuropsychiatric Interview (MINI), Structured Clinical Interview for DSM Disorders (SCID), Composite International Diagnostic Interview (CIDI), Diagnostic Interview Schedule (DIS) or Revised Clinical Interview Schedule (CIS-R)).

Information sources and searches

Two independent reviewers searched Embase, Web of Science, PsycINFO, CINAHL Plus and PubMed from 1998 until June 2015. We used the key words 'PHQ', 'patient health questionnaire', 'screening', 'depression', 'MDD', 'primary care' and 'general practice'.

Data abstraction

We collected information about study characteristics and quality using a standardised data collection form. We included the following characteristics: setting, country, age of sample, gender of sample, year of study, sample size, masking of the assessor of the reference test, data integrity, cut-off score and translation of non-English versions of PHQ-9. When an article appeared to meet the criteria but did not contain sufficient data, we contacted the authors up to two times a month.

Study selection

After the removal of duplicates, two independent reviewers screened the titles and abstracts of all potentially eligible articles. Both authors applied the eligibility criteria, and a list of full text articles was developed through consensus. The two reviewers then considered the full texts of these articles and the final list of included articles was reached through consensus. A third reviewer was available for mediation throughout this process.

Methodological quality assessment

We used the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool to assess risk of bias factors in primary studies, and these factors will be included as study-level variables in analyses.[36] The updated QUADAS-2 guidelines stipulate that it should be adapted for each specific review. We employed the QUADAS-2 adaptation utilised in a recent generic PHQ meta-analysis.[30] The QUADAS-2 incorporates assessments of risk of bias across four core domains: patient selection, the index test, the reference standard and the flow and timing of assessments. Two reviewers independently assessed risk of bias with any discrepancies resolved by consensus. Two reviewers also independently assessed outliers that may be qualitatively different in study design.

Meta-analysis and proposed subgroup analysis

A pooled meta-analysis of suitable studies was conducted to identify overall test accuracy, sensitivity, specificity, combined Youden score, positive and negative predictive values (PPV/NPV), positive and negative likelihood ratios (LR+/LR−) and positive and negative clinical utility index (CUI+/CUI−). Further details are available at www.clinicalutility.co.uk. The CUI is a proxy for the applied value of a test with a qualitative as well as quantitative interpretation.[37–39] Clinical utility may be more important to clinicians than validity.[40] Clinical utility estimates the clinical value of a diagnostic test taking into account both the accuracy of the test and its occurrence. The positive utility index (for rule-in or case-finding accuracy) is a product of sensitivity and PPV and the negative utility index (for rule-out or screening accuracy) is a product of Sp x NPV. The interpretation of the CUI is 0.93–1.00 near perfect value, 0.81–0.92 excellent, 0.64–0.80 good, 0.49–0.63 fair, 0.36–0.48 poor and <0.36 very poor.

Sensitivity and specificity are generally regarded as intrinsic characteristics of a test and independent of prevalence and are a useful initial metric, but these measures do not reflect clinical practice or inform clinicians how to interpret a positive or negative test.[41]

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

Summary measures of diagnostic accuracy typically use receiver operating characteristic (ROC) curve analysis, by which sensitivity and specificity linked with all possible cut-off scores were calculated and plotted.[42] For an individual study, an optimal cut-off score is chosen which balances sensitivity and specificity. ROC curve data are a proportion with a confidence interval which can be combined across all qualifying studies. From the supplied data, we constructed 2 × 2 tables for each cut-off score and computed any missing values. For completeness, we also performed a bivariate meta-analysis to obtain pooled estimates of specificity and sensitivity and their associated 95% confidence intervals (CIs). We constructed summary ROC curves using the bivariate model to produce a 95% confidence ellipse within the ROC curve space. Each data score in this space represents a separate study. We also constructed a Bayesian plot of conditional probabilities which shows all PPVs and NPVs across every possible prevalence.

We assessed between-study heterogeneity using the I^2 statistic[43] which describes the percentage of total variation across studies that is caused by heterogeneity rather than chance. As per convention, we considered an I^2 value of 25% to be low, 50% to be moderate and 75% to be high. We explored the causes of heterogeneity if there was significant between-study heterogeneity. Publication bias was assessed by Harbord or Egger methods.[44]

For a secondary moderator analysis, we performed sub-analysis in clinically relevant subgroups such as those studies with a head-to-head comparison of tools in the same sample. We also attempted a logistic meta-regression analysis of diagnostic accuracy using the 50th percentile of Youden score (sum of sensitivity and specificity) using covariates in the meta-regression model.[45] We investigated heterogeneity resulting from the characteristics of the sample or study design by exploring the effects of potential predictive variables.

Results

Search results

The initial search yielded 777 hits. After removal of duplicates, 621 abstracts and titles were screened (Fig. 1). At the full-text review stage, 58 articles were considered and 32 were subsequently excluded, leaving 26 publications and 40 different analyses that were included in the review.[19,32,33,46–68] Details regarding the search results, including reasons for exclusion of articles are summarised in Fig. 1.

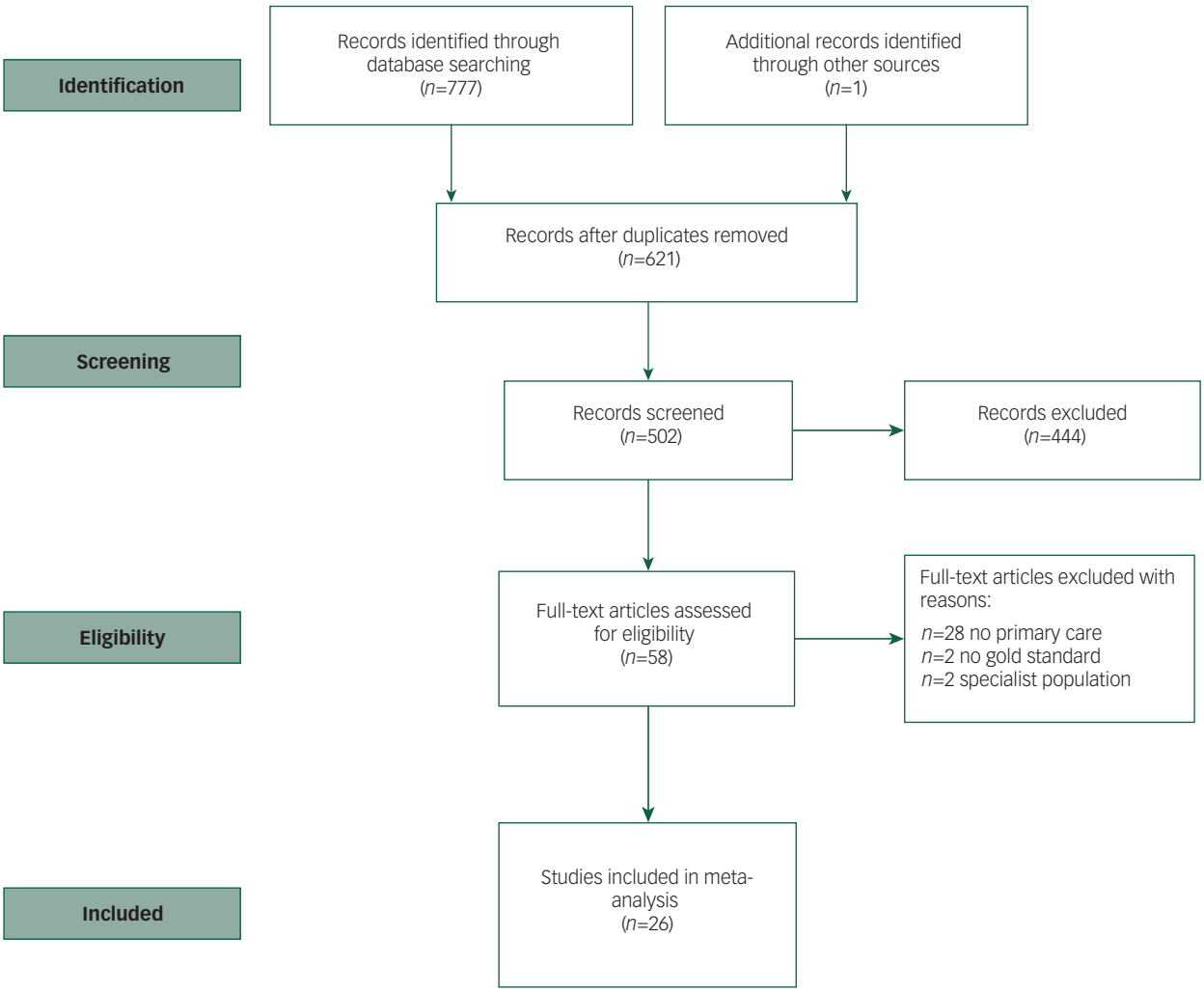

Fig. 1 PRISMA flow diagram of search strategy.

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

Table 1 Summary of included studies

Author	PHQ method	Sample mean age and % male/female	Sample size	Prevalence of depression, %	Reference standard
Lowe *et al*[34]	PHQ-2 (linear) ≥2	42 years, 32.5% male	520	13.7	SCID, DSM-IV
Löwe *et al*[33]	PHQ-9-linear and algorithm	41.7 years, 32.9% male	501	13.2	SCID, DSM-IV
Arroll *et al*[46]	PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10; PHQ-9 (algorithm)	49 years, 39% male	2642	6.2	CIDI, DSM-IV
Ayalon *et al*[47]	PHQ-9 (algorithm)	75 years 59.5% male	153	3.9	SCID, DSM-IV
Azah *et al*[48]	PHQ-9 (linear) ≥10	38.7 years, 38.3% male	180	46.1	CIDI, ICD-10
Cannon *et al*[49]	PHQ-9 (algorithm)	57.2 years, 54% male	526	26.6	SCID DSM-IV MDD lifetime
Chen *et al*[50]	PHQ-9 (linear) ≥10	Age not reported, 47% male	262	16.7	SCID, DSM-IV
Chen *et al*[51]	PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10	68.5 years, 56.3% female	77	54.5	SCID DSM-IV
De Lima Osório *et al*[52]	PHQ-9 (linear) ≥10; PHQ-2 (linear) ≥2	48% under 30, 52% between 31 and 50 years, 100% female	177	33.9	SCID DSM-IV
Gelaye *et al*[53]	PHQ-9 (linear) ≥10	35.1 years, 61.3% female	363	12.7	SCAN DSM-IV
Gilbody *et al*[54]	PHQ-9 (linear) ≥10	42.5 years, 77.1% female	96	37.5	SCID DSM-IV
Henkel *et al*[55]	PHQ-2 (linear) ≥2	53.9 years, 75% female	382	10.0	SCID DSM
Kroenke *et al*[19]	PHQ-9 (linear) ≥10	46 years, 66% female	580	7.1	DSMIIR
Lamers *et al*[56]	PHQ-9 (algorithm); PHQ-9 (linear) ≥8	71.4 years, 51.8% male	713	17.3	MINI DSM-IV
Liu *et al*[57]	PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10	18 years or older, 39% male	1532	3.3	SCAN DSM-IV
Lotrakul *et al*[58]	PHQ-9 (linear) ≥10; PHQ-9 (algorithm)	45 years, 73.7% female	279	6.8	MINI DSM-IV
Patel *et al*[59]	PHQ-9 (linear) ≥10	37.5 years, 56.4% female	598	5.5	ICD-10
Phelan *et al*[60]	PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10	78 years, 62% female	69	11.6	SCID DSM
Richardson *et al*[61]	PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10	15.3 years, 60% female	442	4.3	DIS for MDD in children (DISC)
Sherina *et al*[62]	PHQ-9 (linear) ≥10	30.9 years, 100% female	146	12.1	CIDI, ICD-10
Spitzer *et al*[63]	PHQ-9 (algorithm)	46 years, 66% female	585	10.0	DSM
Sung *et al*[64]	PHQ-9 >6	36.1 years, 65.3% female	400	3.0	MINI DSM
Wittkampf *et al*[66]	PHQ-9 (linear) ≥10; PHQ-9 (algorithm)	49.8 years, 66.7% female	664	12.3	SCID-I
Yeung *et al*[67]	PHQ-9 (linear) ≥15	Not reported	184	22.8	SCID DSM
Zuithoff *et al*[68]	PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10; PHQ-9 (algorithm)	51 years, 37% female	1352	13.0	CIDI DSM-IV

CIDI, Composite International Diagnostic Interview; DIS, Diagnostic Interview Schedule; DISC, Diagnostic Interview Schedule for Children; DSM, Diagnostic and Statistical Manual of Mental Disorders; ICD, International Classification of Diseases; MDD, major depressive disorder; MINI, Mini International Neuropsychiatric Interview; PHQ, Patient Health Questionnaire; SCAN, Schedules for Clinical Assessment in Neuropsychiatry; SCID, Structured Clinical Interview for DSM Disorders.

Study and participant characteristics

Details of the included studies are summarised in Table 1. Briefly, 11 studies examined the PHQ-9, 9 examined the PHQ-9-algoithm and 20 examined the PHQ-9-linear. Several studies compared diagnostic methods within the same population, allowing a head-to-head comparison. Of particular interest, Thompson & Higgens,[45] Manea *et al*[32] and Lowe *et al*[33] compared all three diagnostic methods. Chen *et al*, Kroenke *et al*, Liu *et al*, de Lima Osório *et al*, 2009, Phelan *et al* and Richardson *et al* compared the PHQ-2 with the PHQ-9-linear.[19,50,52,57,60,61] Lamers *et al*, Lotrakul *et al*, Wittkampf *et al* and Zuithoff *et al* compared the PHQ-9-algorithm with the PHQ-9-linear. In these head-to-head studies, the cut-off thresholds were consistent, namely PHQ-2 (linear) ≥2 and PHQ-9 (linear) ≥10.[56,58,66,68]

The total sample size was 26 902 (median 502, s.d.=693.7) with a mean patient age of 49.38 years, and 61% were female. There were 23 706 individuals without depression according to the criterion reference and 3009 with depression, meaning that the prevalence of depression in primary care was 11.3% (95% CI 10.92–11.68%) from simple pooling of data. However, as several publications used multiple tests, after limiting the analysis to unique adults, there were 14 760 people, of whom 2117 had depression (14.3%; 95% CI 11.3–17.7).

Methodological quality

Supplementary Table DS1 summarises the QUADAS-2 scores for all of the included studies. Only four studies were judged low risk of bias across all four domains.[33,45,55,59] Three studies had either high risk of bias or were considered possible outliers. Richardson *et al*,[61] utilised adolescents seen in primary care; Whooley *et al*,[65] used the Whooley questions and was eventually excluded; finally Cannon *et al*,[48] used lifetime risk of depression rather than current depression (although this did not significantly influence the recorded prevalence levels). We used this information as a moderator analysis.

Diagnostic accuracy of the PHQ

Sensitivity and specificity meta-analysis

Main analysis. The diagnostic validity meta-analysis gave overall sensitivity estimates of 82.2% (95% CI 74.3–88.9), 58.4% (95% CI 44.5–71.7) and 89.9% (95% CI 83.4–94.9) for the PHQ-9-linear, PHQ-9-algorithm and PHQ-2 respectively. In all cases, there was significant heterogeneity but no significant publication

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

Table 2 Inconsistency and bias analysis in Patient Health Questionnaire (PHQ) data in primary care

Test	Sensitivity bias, % (95% CI)	Specificity bias, % (95% CI)	ROC bias, % (95% CI)
Main results			
PHQ-9-linear *n*=20	I^2 (inconsistency) = 90.4 (87 to 92.5) Harbord: bias = 3.08 (92.5% CI −0.604 to 6.77) *P*=0.1314	I^2 (inconsistency) = 96.6 (96 to 97.1) Harbord: bias = −5.742 (92.5% CI −10.38 to −1.097) ***P*=0.0312**	I^2 (inconsistency) = 77.2 56.8 to 85.6) Egger: bias = −0.805 (−2.50 to 0.89) *P*=0.3154
PHQ-9-algorithm *n*=9	I^2 (inconsistency) = 94.8 (92.7 to 96.1) Harbord: bias = 0.363 (92.5% CI −7.707 to 8.433) *P*=0.9277	I^2 (inconsistency) = 98.3 (97.9 to 98.5) Harbord: bias = −13.19 (92.5% CI −31.689 to 5.302) *P*=0.1797	I^2 (inconsistency) = 92.1 (87.8 to 94.4) Egger: bias = 1.33 (−7.685 to 10.345) *P*=0.7375
PHQ-2 *n*=11	I^2 (inconsistency) = 85.3 (74.9 to 90.2) Harbord: bias = 0.98 (92.5% CI −2.846 to 4.815) *P*=0.6175	I^2 (inconsistency) = 97 (96.2 to 97.5) Harbord: bias = −3.89 (92.5% CI −11.382 to 3.593) *P*=0.3225	I^2 (inconsistency) = 73.2 (14 to 86.4) Egger: bias = −0.64 (−6.848 to 5.554) *P*=0.7865
Head-to-head results			
PHQ-9-linear *n*=8	I^2 (inconsistency) = 88.5 (79.4 to 92.5) Harbord: bias = 1.22693 (92.5% CI −4.335 to 6.789) *P*=0.6519	I^2 (inconsistency) = 96.7 (95.6 to 97.4%) Harbord: bias = −4.17 (92.5% CI −12.913 to 4.558) *P*=0.3433	I^2 (inconsistency) = 86.2 (57.6 to 92.8) Egger: bias = −1.03 (−11.014 to 8.950) *P*=0.6999
PHQ-2-linear *n*=8	I^2 (inconsistency) = 83.2 (65.5 to 89.8) Harbord: bias = −0.18 (92.5% CI −4.559 to 4.198) *P*=0.9321	I^2 (inconsistency) = 97.1 (96.2 to 97.7) Harbord: bias = −4.01 (92.5% CI −12.868 to 4.830) *P*=0.3664	I^2 (inconsistency) = 17.7 (0 to 73.2) Egger: bias = −1.93 (−4.395 to 0.526) *P*=0.0774
PHQ-9-algorithm *n*=6	I^2 (inconsistency) = 94.2 (90.5 to 96) Harbord: bias = 6.96 (92.5% CI −4.920 to 18.852) *P*=0.2336	I^2 (inconsistency) = 91.2 (83.7 to 94.3) Harbord: bias = −5.02 (92.5% CI −15.119 to 5.069) *P*=0.2996	I^2 (inconsistency) = 92.4 (83.6 to 95.5) Egger: bias = 3.27 (−25.876 to 32.416) *P*=0.6769
Moderator analysis			
PHQ-2 *n*=9 Higher quality; same cut-off; adults	I^2 (inconsistency) = 87.1 (77.1 to 91.5) Harbord: bias = 1.22 (92.5% CI −3.194 to 5.641) *P*=0.5809	I^2 (inconsistency) = 95.8 (94.2 to 96.7) Harbord: bias = −1.86 (92.5% CI −9.244 to 5.506) *P*=0.6128	I^2 (inconsistency) = 76.3 (0 to 89.4) Egger: bias = −0.44 (−13.793 to 12.893) *P*=0.8979
PHQ-9-linear *n*=16 Higher quality; same cut-off; adults	I^2 (inconsistency) = 92.1 (89.3 to 93.9) Harbord: bias = 3.46 (92.5% CI −1.240 to 8.162) *P*=0.1786	I^2 (inconsistency) = 95.9% (94.9 to 96.6) Harbord: bias = −5.86 (92.5% CI −10.423 to 1.313) ***P*=0.0266**	I^2 (inconsistency) = 80.8 (61.1 to 88.3) Egger: bias = −3.20 (−6.581 to 0.173) *P*=0.0598
PHQ-9-algorithm *n*=8 Higher quality; same cut-off; adults	I^2 (inconsistency) = 95.1 (93 to 96.4) Harbord: bias = 0.35 (92.5% CI −8.228 to 8.942) *P*=0.9316	I^2 (inconsistency) = 98.1 (97.6 to 98.4) Harbord: bias = −12.03 (92.5% CI −33.868 to 9.808) *P*=0.2808	I^2 (inconsistency) = 93 (89.1 to 95) Egger: bias = 1.41 (−8.589 to 11.426) *P*=0.7406

ROC, receiver operating characteristic.
Values in bold are significant.

bias (see Table 2 which contains the heterogeneity and publication bias data for all of the pooled analysis). The pooled specificity was 84.7% (95% CI 80.4–88.5), 92.1% (95% CI 85.9–96.6) and 72.6% (95% CI 66.0–78.7) for the PHQ-9-linear, PHQ-9-algorithm and PHQ-2 respectively. In the sensitivity analysis (in which we removed the three outliers) and in the bivariate analysis, the results were broadly unchanged (Table 3 and Fig. 2) but they did generate our best estimate of sensitivity of 81.3% (95% CI 71.6–89.3) and specificity of 85.3% (95% CI 81.0–89.1) for the PHQ-9-linear; a best estimate of sensitivity of 89.3% (95% CI 81.5–95.1) and specificity of 75.9% (95% CI 70.1–81.3) for the PHQ-2; a best estimate of sensitivity of 56.8% (95% CI 41.2–71.8) and specificity of 93.3% (95% CI 87.5–97.3) for the PHQ-9-algorithm.

Subanalysis (head to head) PHQ-9-linear *v.* PHQ-2. In a subanalysis restricted to head-to-head studies on the same population, the sensitivity of the PHQ-9-linear was 87.0% (95% CI 75.81–95.07) *v.* 91.4% (95% CI 83.60–96.92) for the PHQ-2. The specificity of the PHQ-9-linear was 87.17 (95% CI 81.10–92.20) *v.* 72.23% (95% CI 63.96–79.81) for the PHQ-2. In the sensitivity analysis, the results were unchanged (Table 3).

Subanalysis (head to head) PHQ-9-linear *v.* PHQ-9-algorithm. In a subanalysis restricted to head-to-head studies on the same population, the sensitivity of the PHQ-9-linear was 81.1% (95% CI 63.34–93.86) *v.* 53.1% (95% CI 36.44–69.31) for the PHQ-9-algorithm. The specificity of the PHQ-9-linear was 86.34% (95% CI 80.36–91.38) *v.* 95.71% (95% CI 93.54–97.45) for the PHQ-9-algorithm, suggesting significantly lower specificity for the PHQ-9-linear. However, caution is necessary as these results are from a predefined cut-point of >10. Results were broadly unchanged in the sensitivity analyses (Table 3).

Cut-off analysis: effect of cut-off thresholds. In an analysis restricted to specific cut-offs, we analysed the effect of choosing different fixed cut-off thresholds on the PHQ-2 and PHQ-9 when scored using linear scoring. Results are shown in Table 3. Inevitably, as the cut-point increased sensitivity reduced and specificity increased. For the PHQ-9, looking at combined sensitivity and specificity (Youden index), the optimal cut-off would be ≥10 followed by ≥11.

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

Table 3 Summary of Patient Health Questionnaire (PHQ) analysis in primary care

Test	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV	NPV	ROC	CUI+	CUI-	LR+	LR-
Main results									
PHQ-9-linear *n*=20	82.2 (74.3–88.9)	84.7 (0.80–0.89)	38.0% (36.1–39.9%)	97.7% (97.3–98.0%)	0.910 (0.892–0.930)	0.312 (0.311–0.313) 'very poor'	0.827 (0.826–0.828) 'excellent'	5.37 (5.09–5.67)	0.21 (0.19–0.24)
PHQ-9-algorithm *n*=9	58.4 (44.5–71.7)	92.1 (85.9–96.6)	47.4% (44.7–50.1%)	94.8% (94.3–95.3%)	0.733 (0.676–0.795)	0.277 (0.276–0.278) 'very poor'	0.873 (0.873–0.873) 'excellent'	7.39 (6.77–8.07)	0.45 (0.42–0.49)
PHQ-2 *n*=11	89.9 (83.4–94.9)	72.6 (66.0–78.7)	23.1% (21.5–24.7%)	97.5% (97.0–97.9%)	0.860 (0.819–0.903)	0.188 (0.187–0.189) 'very poor'	0.708 (0.707–0.709 'good'	2.97 (2.82–3.12)	0.26 (0.22–0.30)
Head-to-head results									
PHQ-9-linear *n*=8	87.0 (75.8–95.1)	87.2 (81.1–92.2)	38.2% (35.1–41.2%)	98.6% (98.3–98.9%)	0.920 (0.915–0.924)	0.322 (0.321–0.323)'very poor'	0.877 (0.987–0.878) 'excellent'	7.66 (7.04–8.33)	0.18 (0.14–0.22)
PHQ-2-linear *n*=8	91.5 (83.6–96.9)	72.2 (64.0–79.8)	22.7% (20.8–24.6%)	99.0% (98.7–99.3%)	0.900 (0.865–0.934)	0.205 (0.205–0.206) 'very poor'	0.742 (0.741–0.743 'good'	3.61 (3.42–3.81)	0.13 (0.10–0.17)
PHQ-9-algo *n*=6	53.0 (36.4–69.3)	95.7 (93.5–97.5)	58.4% (54.1–62.8%)	94.1% (93.5–94.7%)	0.715 (0.628–0.815)	0.273 (0.272–0.275) 'very poor'	0.905 (0.904–0.906) 'excellent'	12.35 (10.56–14.45)	0.55 (0.51–0.50)
Moderator analysis									
PHQ-9-linear *n*=16 Higher quality; same cut-off; adults	81.3 (71.6–89.3) 82.33(72.0–89.4)[a]	85.3 (81.0–89.1) 86.4 (81.2–90.4)[a]	44.2% (41.9–46.6%)	38.9 (36.8–41.0)	97.5 (97.2–97.9)	0.316 (0.315–0.317) 'very poor'	0.832 (0.832–0.832) 'excellent'	5.53 (5.21–5.87)	0.22 (0.19–0.25)
PHQ-9-algorithm *n*=8 Higher quality; same cut-off; adults	56.8 (41.2–71.8) 54.0 (40.0–67.5)[a]	93.3 (87.5–97.3) 95.9 (94.0 – 97.3)[a]	60.3% (57.0–63.6%)	48.3% (45.4–51.3)	95.1% (94.7–95.6)	0.275 (0.274–0.276) 'very poor'	0.887 (0.887–0.887) 'excellent'	8.46 (97.67 –9.33)	0.46 (0.43–0.50
PHQ-2 *n*=9 Higher quality; same cut-off; adults	89.3 (81.5–95.1) 91.4 (81.9–96.2)[a]	75.9% (70.1–81.3) 76.3 (69.6–82.0)[a]	27.7% (25.8–29.6%)	26.5% (24.6–28.3%)	98.6% (98.3–99.0%)	0.236 (0.235–0.237) 'very poor'	0.749 (0.749–0.749) 'good'	3.71 (3.52–3.90	0.14 (0.11–0.18)

PPV, positive predictive value; NPV, negative predictive value; ROC, receiver operating characteristic; CUI+, positive clinical utility index; CUI–, negative clinical utility index; LI+, positive likelihood ratio; LI–, negative likelihood ratio.
a. Alternative calculation based on bivariate calculation in STATA.

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

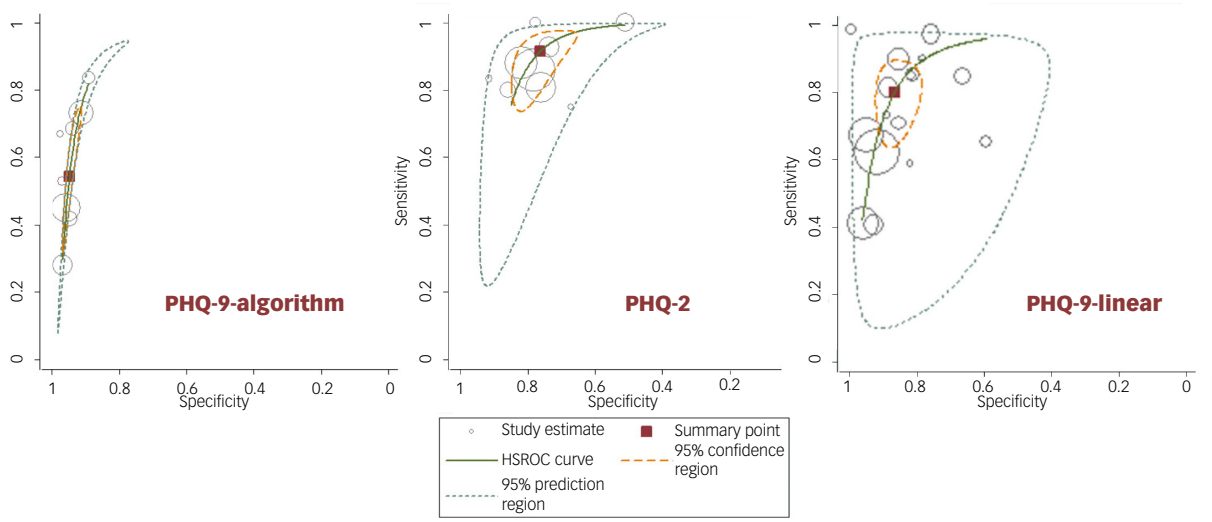

Fig. 2 Bayesian plot of conditional probabilities PHQ-9-linear *v*. PHQ-9-algoithm *v*. PHQ-2 (restricted to head-to-head studies).

Moderator analysis: effect of influencing variables. In a moderator analysis we found no association between country, mean age, gender, year of publication or sample size.

ROC curve meta-analysis

Main analysis PHQ-9 linear, PHQ algorithm and PHQ-2. The pooled ROC diagnostic validity meta-analysis gave an overall area estimate of 0.91 (95% CI 0.892–0.930) for the PHQ-9-linear, 0.733 (95% CI 0.676–0.795) for the PHQ-9-algorithm and 0.860 (95% CI 0.819–0.903) for the PHQ-2. In all cases there was significant heterogeneity but no significant publication bias; see Table 3 (summary of results). Results were broadly unchanged in moderator analysis with area under the ROC of 0.910 (95% CI 0.882–0.939) for the PHQ-9 linear, 0.732 (0.667–0.803) for the PHQ-9-algorithm and 0.877 (0.824–0.934) for the PHQ-2.

Subanalysis (head to head) PHQ-9-linear *v*. PHQ-2. The area under the ROC for the PHQ-2 was 0.898 (95% CI 0.864–0.933) and 0.922 (95% CI 0.882–0.964) for the PHQ-9-linear in the head-to-head studies. Once again, results were unchanged in the sensitivity analysis.

Subanalysis (head to head) PHQ-9-linear *v*. PHQ-9-algorithm. The area under the ROC for the PHQ-9-linear was 92.01 (95% CI 91.53–92.48) and 71.49 (95% CI 62.75–81.45) for the PHQ-9-algorithm when restricted to four head-to-head studies.

Subanalysis (head to head) PHQ-2 *v*. PHQ-9-algorithm. There were insufficient data for this comparison.

Test performance: case finding *v*. screening

Examining PPV, the diagnostic validity meta-analysis suggested superior PPV of the PHQ-9-algorithm 47.4% (95% CI 44.7–50.1) compared with the PHQ-2 23.1% (95% CI 21.5–24.7); however, caution is required because prevalence is not controlled for (i.e. not matched in both analyses) (correction for prevalence is shown in the Bayesian curve of conditional probabilities). Examining NPV, meta-analysis suggested superior PPV of the PHQ-2 97.5% (95% CI 97.0–97.9) compared with the PHQ-9-algorithm 94.8% (95% CI 94.3–95.3); however, caution is again required because prevalence is not controlled for in this analysis. Results using likelihood ratios are shown in Table 3 but more informative is the clinical utility. For case finding (CUI+), all methods were disappointing, with the following results: PHQ-9-linear 0.312 (95% CI 0.311–0.313), PHQ-9-algorithm 0.277 (95% CI 0.276–0.278) and PHQ-2 0.188 (95% CI 0.187–0.189), all suggesting very poor performance at typical prevalence rates seen in primary care. Results were not substantially different using a moderator analysis for high-quality studies with a fixed cut-off or using head-to-head analysis.

For application as a screening test (CUI+) all methods were satisfactory with the following results: PHQ-9-linear 0.827 (95% CI 0.826–0.828), PHQ-9-algorithm 0.873 (95% CI 0.873–0.873) and PHQ-2 0.708 (95% CI 0.707–0.709), all suggesting good to excellent performance at typical prevalence rates seen in primary care. Results were not substantially different using a moderator analysis for high-quality studies with a fixed cut-off or using head-to-head analysis. All analyses suggested the optimal rule-out screening test would be the PHQ-9-algorithm, closely followed by the PHQ linear.

Using a Bayesian curve of conditional probabilities, the performance of each test (judged by PPV and NPV) can be demonstrated at every possible prevalence applicable to different settings (Figs. 3 and 4). From the Bayesian curve, the most encouraging test would be the PHQ-2 used as an initial screener followed by either the PHQ-9-linear or another suitable case-finding tool.

Cut-off analysis: effect of cut-off thresholds

On the PHQ-9, looking at combined PPV and NPV (predictive summary index), the optimal cut-point would be ≥14. For the PHQ-2, looking at combined sensitivity and specificity (Youden index), the optimal cut-off would be ≥3 closely followed by ≥2 (note that ≥2 is the conventional threshold). However, looking at combined PPV and NPV (predictive summary index), the optimal cut-point would be ≥6, followed by the ≥5. Comparing the PHQ-9 and the PHQ-2 across all possible cut-offs shows that neither is satisfactory as a case-finding tool in primary care at any cut-off, but the optimal single method is the PHQ-2 at a threshold of ≥6. Six per cent of those without MDD have a score of 5 or lower on the PHQ-2 and of those with a score of 5 or lower, 93.5% are true negatives (true non-cases) (Tables 2–4).

Discussion

A previous meta-analysis of 41 studies involving 50 371 individuals in primary care found a pooled prevalence of 18.4% (95% CI 13.5–23.9) in adults aged 18–65 years using semi-structured

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

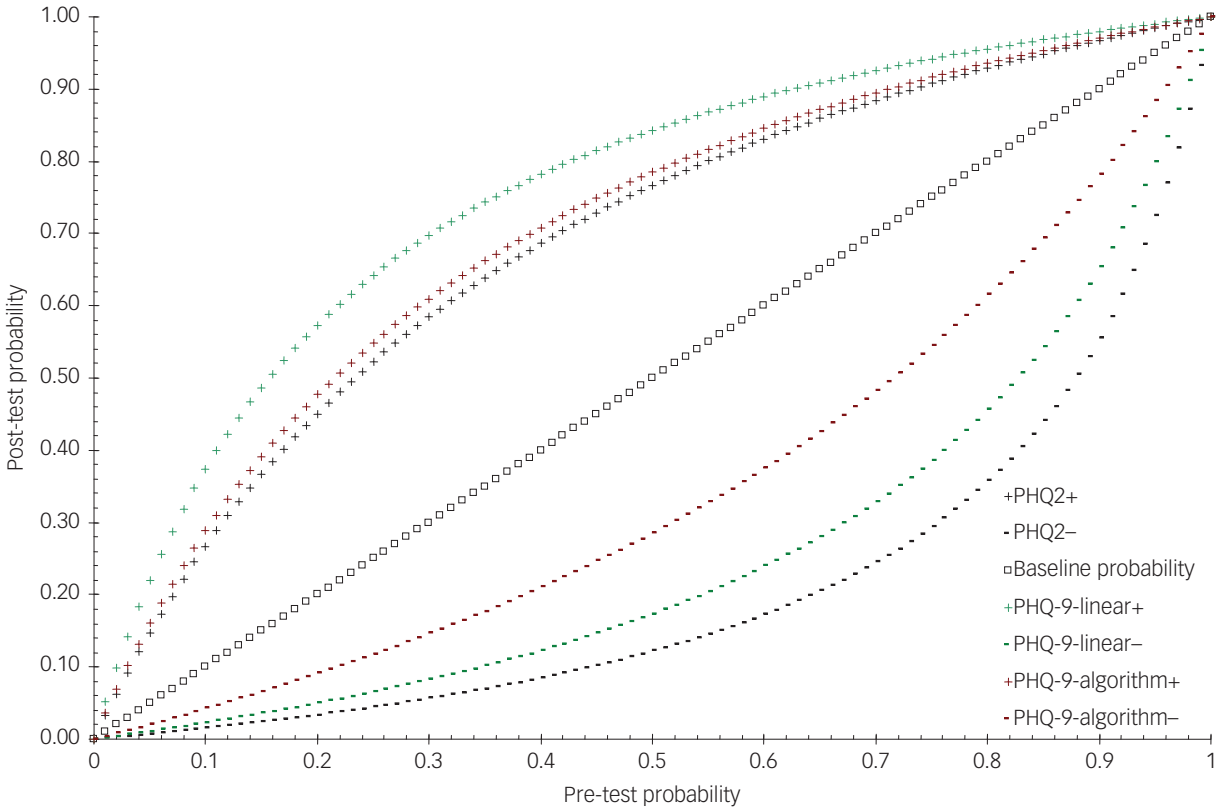

Fig. 3 Bivariate plot of summary accuracy of PHQ-9-linear *v*. PHQ-9-algorithm *v*. PHQ-2 (restricted to high-quality studies).

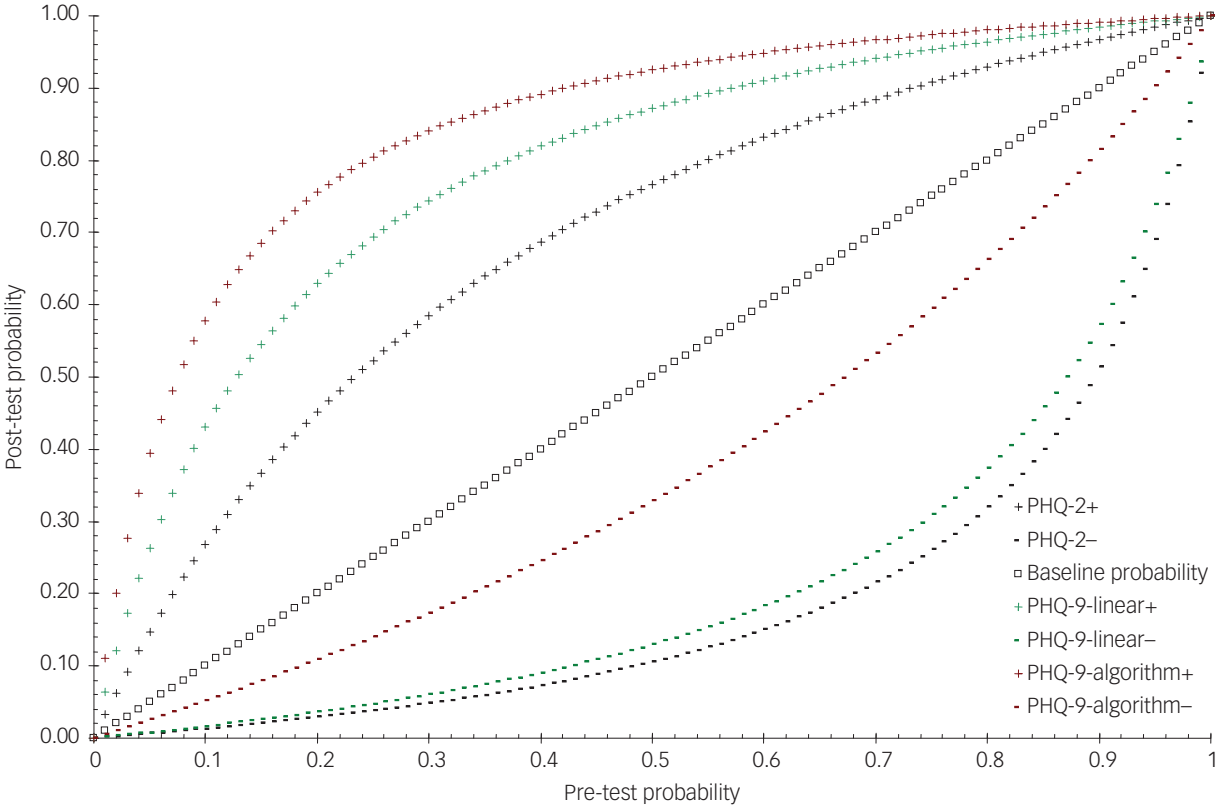

Fig. 4 Bayesian plot of conditional probabilities PHQ-9-linear *v*. PHQ-9-algoithm *v*. PHQ-2.

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

Table 4 PHQ cut-off threshold analysis in primary care

Test	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV	NPV	CUI+	CUI–	LR+	LR–
PHQ-2								
Cut PHQ-2 ≥1 *n*=20	96.05 (92.29–98.60)	52.18 (43.42–60.8)	14.7 (13.6–15.9)	99.4 (99.1–99.6)	0.141 (0.140–0.142) 'very poor'	0.519 (0.519–0.519) 'fair'	2.01 (1.95–2.07)	0.07 (0.05–0.11)
Cut PHQ-2 ≥2 *n*=9	92.20 (85.21–97.10)	70.98 (64.63–76.94)	22.8 (21.2–24.5)	99.0 (98.7–99.3)	0.211 (0.210–0.212) 'very poor'	0.703 (0.703–.703) 'good'	3.18 (3.04–3.32)	0.11 (0.08–0.14)
Cut PHQ-2 ≥3 *n*=11	76.22 (61.1–88.53)	88.66 (85.01–91.86)	38.6 (35.9–41.3)	97.6 (97.2–97.9)	0.294 (0.293–0.295) 'very poor'	0.865 (0.865–0.865) 'excellent'	6.74 (6.23–7.30)	0.27 (0.23–0.31)
Cut PHQ-2 ≥4 *n*=11	61.46 (44.00–77.52)	94.14 (91.73–96.15)	53.4 (49.7–57.2)	95.7 (95.2–96.3)	0.329 (0.328–0.330 'very poor'	0.901 (0.901–0.901) 'excellent'	10.45 (9.22–11.85)	0.41 (0.37–0.45)
Cut PHQ-2 ≥5 *n*=11	47.33 (26.78–68.37)	97.60 (95.36–99.12)	72.8 (67.2–78.4)	93.2 (92.2–94.1)	0.344 (0.241–0.346) 'very poor'	0.909 (0.908–0.910) 'excellent'	19.77 (15.23–25.68)	0.54 (0.49–0.60)
Cut PHQ-2 ≥6 *n*=11	51.80 (23.07–79.88)	98.63 (96.91–99.66)	83.3 (78.5–88.2)	93.5 (92.6–94.4)	0.421 (0.419–0.424) 'poor'	0.421 (0.419–0.424)[b] 'poor'	0.922 (0.921–0.923)[a] 'excellent'	0.50 (0.45–0.56)
PHQ-9								
Cut PHQ-9 ≥6 *n*=20	89.81 (81.91–95.63)	62.79 (51.02–73.84)	28.9 (26.7–31.1)	97.3 (96.6–98.0)	0.259 (0.258–0.250) 'very poor'	0.611 (0.611–0.611) 'fair'	2.41 (2.28–2.55)	0.16 (0.13–0.21)
Cut PHQ-9 ≥7 *n*=20	84.69 (74.32–92.75)	69.17 (57.72–79.53)	31.6 (29.2–34.1)	96.4 (95.6–97.2)	0.268 (0.267–0.269) 'very poor'	0.667 (0.666–0.668) 'good'	2.75 (2.57–2.93)	0.22 (0.18–0.27)
Cut PHQ-9 ≥8 *n*=20	80.25 (71.00–88.09)	76.54 (69.59–82.84)	29.4 (27.4–31.5)	96.9 (96.4–97.5)	0.236 (0.235–0.237) 'very poor'	0.742 (0.71–0.743) 'good'	3.41 (3.21–3.63)	0.26 (0.22–0.30)
Cut PHQ-9 ≥9 *n*=20	81.31 (69.69–90.64)	79.82 (72.51–86.26)	32.3 (30.0–34.7)	97.3 (96.8–97.8)	0.263 (0.262–0.264) 'very poor'	0.777 (0.776–0.778) 'good'	4.03 (3.77–4.31)	0.23 (0.20–0.28)
Cut PHQ-9 ≥10 *n*=20	81.3 (71.6–89.3)	85.3 (81.0–89.1)	44.2 (41.9–46.6)	97.0 (96.7–97.4)	0.333 (0.337–0.339) 'very poor'	0.863 (0.863–0.863) 'excellent'	6.90 (6.44–7.40)	0.27 (0.24–0.30)
Cut PHQ-9 ≥11 *n*=20	75.40 (60.77–87.52)	87.86 (82.77–92.17)	44.6 (41.8–47.4)	96.5 (96.0–97.0)	0.336 (0.335–0.337) 'very poor'	0.848 (0.848–0;.848) 'excellent'	6.23 (5.74–6.76)	0.28 (0.25–0.32)
Cut PHQ-9 ≥12 *n*=20	68.37 (54.71–80.58)	90.88 (87.54–93.73)	49.1 (46.1–52.0)	95.7 (95.2–96.2)	0.336 (0.335–0.336) 'very poor'	0.870 (0.870–8.70) 'excellent'	7.51 (6.85–8.23)	0.35 (0.31–0.39)
Cut PHQ-9 ≥13 *n*=20	69.92 (58.39–80.30)	92.93 (89.33–95.83)	60.2 (55.5–64.9)	95.3 (94.4–96.1)	0.421 (0.419–0.423) 'poor'	0.421 (0.419–0.423)[b] 'poor'	9.84 (8.38–11.55)	0.32 (0.28–0.38)
Cut PHQ-9 ≥14 *n*=20	56.04 (42.88–68.77)	96.57 (94.48–98.18)	73.4 (67.1–79.8)	92.9 (91.6–94.2)	0.411 (0.408–0.415) 'poor'	0.898 (0.897–0.898)[a] 'excellent'	16.5 (12.26–22.21)	0.46 (0.40–0.53)

PPV, positive predictive value; NPV, negative predictive value; ROC, receiver operating characteristic; CUI+, positive clinical utility index; CUI–, negative clinical utility index; LI+, positive likelihood ratio; LI–, negative likelihood ratio.
a. Optimal cut-off for ruling out those without depression (screening).
b. Optimal cut-off for ruling in those with depression (case-finding).

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

interviews.[1] In this study, we found a slightly lower prevalence of depression in primary care of 14.3% (95% CI 11.3–17.7%) across 14 760 adults. The PHQ-9-linear had better sensitivity but worse specificity than the PHQ-9-algorithm. However, this finding could result from choosing a PHQ-9-linear cut-off threshold which is too low. Regarding the PHQ-2, it had significantly greater specificity over the PHQ-9-linear method. Analysis using the ROC meta-analysis suggested that the area under the ROC of the PHQ-9-linear as well as that of the PHQ-2 was significantly higher than the PHQ-9-algorithm which was surprising given that the PHQ-9-algorithm more tightly adheres to the DSM criterion standard. The difference was maintained when PHQ-9-linear and PHQ-9-algorithm were compared with analysis was restricted to four head-to-head studies. In head-to-head studies, the tools are tested against one another in the same sample, ruling out differences according to prevalence or local conditions. Using the same methods, there was no clear differences between the PHQ-2 and PHQ9-linear which again is surprising given the brevity of the PHQ-2.

However, these results do not clarify a specific role for any method in either screening or case finding. For case finding, consistent with previous literature, all methods were disappointing with the results on the CUI+ graded as 'very poor'. Looking at PPV alone for all methods using the Bayesian curve, results were similarly poor thus confirming overall poor performance of this method at typical prevalence rates seen in primary care. In short, a positive test is infrequent in a typical primary care sample and/or a positive test (when it does occur) is not especially discriminating. For application as a screening test all methods were encouraging with the following results on the CUI−: PHQ-9-linear 0.827 (0.826–0.828), PHQ-9-algorithm 0.873 (0.873–0.873) and PHQ-2 0.708 (0.707–0.709), all suggesting good to excellent performance at typical prevalence rates. In NPV, values were all high. Examining this effect in more detail using a Bayesian curve of conditional probabilities demonstrated (Figs 3 and 4) that although none of the methods performed particularly well at case finding at any prevalence rate when used alone, they performed reasonably well at initial first step. The most practical use of these tools would be the PHQ-2 used as an initial screener followed by either the PHQ-9-linear or another suitable case-finding tool.

We also analysed the effect of varying the cut-point. If simply considering sensitivity and specificity, then the cut-point analysis suggested that the current thresholds of ≥10 on the PHQ-9 and ≥2 on the PHQ-2 are very close to optimal. However, as discussed above there is more to the application of tests in clinical practice than simply looking at combined sensitivity and specificity. Clinical utility is better represented by PPV and NPV. Using PPV and NPV (combined) suggests that a substantially higher cut-point in both the PHQ-9 and the PHQ-2 may be appropriate. Furthermore, if one discounts their role in case finding and simply concentrates on rule-out ability (CUI–), then the optimal cut points would be ≥14 on the PHQ-9 and ≥6 on the PHQ-2. Although these high thresholds are surprising, it is evident that those without MDD, 98.6% have a score of 5 or lower on the PHQ-2 and of those with a score of 5 or lower, 93.5% are true negatives (true non-cases). Similarly, 96.5% of non-cases scored <14 on the PHQ-9 and of those that do, 92.9% are true negatives. We suggest further work is required to examine the optimal cut-off thresholds if a two-step procedure were to be used.

Limitations

We acknowledge that there were relatively few studies with all the required subgroups and not all studies reported ROC data (but we were able to calculate this in many cases). To date, studies have not attempted to clarify whether the sample comprises previously untreated or previously undiagnosed patients. We did not attempt to look at severity assessment or sensitivity to change. It must also be acknowledged that the results presented represent the outcome of a single application of the PHQ. Multiple (serial) applications may be conducted in clinical practice and would change results. For completeness, if the PHQ-2 is initially applied (step 1), followed by PHQ-9-linear to those who score positive in step 1, then the combined sensitivity would be 72.4% and specificity 96.4% (overall accuracy 93.0%). If the PHQ-2 were to be initially applied followed by PHQ-9-algorithm, the combined sensitivity would be 50.7% and specificity 98.4% (overall accuracy 91.6%).

Clinical implications and further research

The PHQ has potential to be used to rule out those without depression with few false negatives but an adjustment of the cut-off points (≥14 on the PHQ-9 and ≥6 on the PHQ-2) should be considered. Alternatively its routine use can be improved by a two-step procedure using PHQ-2 and then PHQ-9. This would also reduce the burden on clinicians as the PHQ-9 would only be applied following a positive initial PHQ-2 screen. Depression tools applied for the purpose of screening and/or case finding will only be of use if combined with adequate follow-up and adequate treatment. Screening without removal of barriers to high-quality care is potentially frustrating and arguably counterproductive. Several reviews found modest evidence to support QoF-based PHQ scoring in part because primary care clinicians may lack the skills or resources to appropriately follow-up a positive screen.[26,69] Further work on cut-off thresholds and repeat assessment may further improve results but care must be taken not to increase the burden on clinicians if they are required to implement screening tools.

This meta-analysis confirms that neither the PHQ-9 nor the PHQ-2 can confirm a diagnosis of MDD when used alone as a one-off measure and this is independent of the scoring method. However, the PHQ-9 and indeed the PHQ-2 can be used as an initial first screening step and indeed performs quite well in this regard.

Alex J. Mitchell, MD, Department of Cancer Studies, University of Leicester, and Department of Psycho-Oncology, Leicestershire Partnership NHS Trust, Leicester, UK; **Motahare Yadegarfar**, MBChB, Medical School, University of Leicester, Leicester, UK; **John Gill**, MBChB, Medical School, University of Leicester, Leicester, UK; **Brendon Stubbs**, PhD, Institute of Psychiatry, Psychology and Neuroscience, King's College London, and Physiotherapy Department, South London and Maudsley NHS Foundation Trust, UK

Correspondence: Alex J. Mitchell, Psycho-Oncology, Department of Cancer Studies, University of Leicester, Leicester LE1 5WW, UK. Email: ajm80@le.ac.uk

First received 3 Jul 2015, final revision 17 Dec 2015, accepted 21 Dec 2015

Acknowledgements

We thank Jemma Adams for her help during the revision of this manuscript.

References

1 Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet* 2009; **374**: 609–19.

2 National Collaborating Centre for Mental Health. *Depression in Adults with a Chronic Physical Health Problem: The NICE Guideline on Treatment and Management* 2010. British Psychological Society & Royal College of Psychiatrists.

3 Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B.Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet* 2007; **370**: 851–8.

4 Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *Lancet* 2013; **382**: 1575–86.

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

5 Harman JS, Veazie PJ, Lyness JM. Primary care physician office visits for depression by older Americans. *J Gen Intern Med* 2006; **21**: 926–30.

6 Shah A. The burden of psychiatric disorder in primary care. *Int Rev Psychiatry* 1992; **4**: 243–50.

7 Zastrow A, Faude V, Seyboth F, Niehoff D, Herzog W, Löwe B. Risk factors of symptom underestimation by physicians. *J Psychosom Res* 2008; **64**: 543–51.

8 Duhoux A, Fournier L, Gauvin L, Roberge P. Quality of care for major depression and its determinants: a multilevel analysis. *BMC Psychiatry* 2012; **12**: 142.

9 Druss BG, Wang PS, Sampson NA, Olfson M, Pincus HA, Wells KB, et al. Understanding mental health treatment in persons without mental diagnoses: results from the National Comorbidity Survey replication. *Arch Gen Psychiatry* 2007; **64**: 1196–203.

10 Takayanagi Y, Spira A, Bienvenu O, Hock RS, Carras MC, Eaton WW, et al. Antidepressant use and lifetime history of mental disorders in a community sample: results from the Baltimore Epidemiologic Catchment Area Study. *J Clin Psychiatry* 2015; **76**: 40–4.

11 Duhoux A, Fournier L, Menear M. Quality indicators for depression treatment in primary care: a systematic literature review. *Curr Psychiatry Rev* 2011; **7**: 104–37.

12 Mojtabai R. Clinician-identified depression in community settings: concordance with structured-interview diagnoses. *Psychother Psychosom* 2013; **82**: 161–9.

13 Dowrick C, Frances A. Medicalising unhappiness: new classification of depression risks more patients being put on drug treatment from which they will not benefit. *BMJ* 2013; **347**: 7140.

14 Jerant A, Kravitz RL, Fernandez Y, Garcia E, Feldman MD, Cipri C, et al. Potential antidepressant overtreatment associated with office use of brief depression symptom measures. *J Am Board Fam Med* 2014; **27**: 611–20.

15 Mitchell AJ, Meader N, Davies E, Clover K, Carter GL, Loscalzo MJ, et al. Meta-analysis of screening and case finding tools for depression in cancer: evidence based recommendations for clinical practice on behalf of the Depression in Cancer Care consensus group. *J Affect Disord* 2012; **140**: 149–60.

16 National Institute for Health and Social Care Excellence. *About the Quality and Outcomes Framework (QOF).* NICE, 2013 (http://www.thementalelf.net/mental-health-conditions/depression/lack-of-evidence-to-support-qof-incentives-for-assessing-depression-severity-using-tools-in-primary-care/#sthash.5RRS2Lm5.dpuf).

17 Mitchell AJ, Coyne JC. Do ultra-short screening instruments accurately detect depression in primary care? A pooled analysis and meta-analysis of 22 studies. *Br J Gen Pract* 2007; **57**: 144–51.

18 Mitchell AJ, Vahabzadeh A, Magruder K. Screening for distress and depression in cancer settings: 10 lessons from 40 years of primary-care research. *Psychooncology* 2011; **20**: 572–84.

19 Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; **16**: 606–13.

20 Kroenke K, Spitzer RL, Williams JB. The patient health questionnaire-2: validity of a two-item depression screener. *Med Care* 2003; **41**: 1284–92.

21 MacMillan HL, Patterson CJ, Wathen CN, Feightner JW, Bessette P, Elford RW, et al. Canadian Task Force on Preventive Health Care: screening for depression in primary care: recommendation statement from the Canadian Task Force on Preventive Health Care. *CMAJ* 2005; **172**: 33–5.

22 U.S. Preventive Services Task Force. Screening for depression: recommendations and rationale. *Ann Intern Med* 2002; **136**: 760–4.

23 Siu AL; US Preventive Services Task Force. Screening for depression in adults: US Preventive Services Task Force Recommendation Statement. *JAMA* 2016; **315**: 380–7.

24 National Collaborating Center for Mental Health. *The NICE Guideline on The Management and Treatment of Depression in Adults (Updated Edition).* National Institute for Health and Clinical Excellence, 2010.

25 Allaby M. *Screening for Depression: A Report for the UK National Screening Committee (Revised Report).* UK National Screening Committee, 2010.

26 Shaw EJ, Sutcliffe D, Lacey T, Stokes T. Assessing depression severity using the UK Quality and Outcomes Framework depression indicators: a systematic review. *Br J Gen Pract* 2013; **63**: e309–17.

27 Joffres M, Jaramillo A, Dickinson J, Lewin G, Pottie K, Shaw E, et al. Canadian Task Force on Preventive Health Care: recommendations on screening for depression in adults. *CMAJ* 2013; **185**: 775–82.

28 Goodyear-Smith FA, van Driel ML, Arroll B, Del Mar C. Analysis of decisions made in meta-analyses of depression screening and the risk of confirmation bias: acase study. *BMC Med Res Methodol* 2012; **12**: 76.

29 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**: 1596–602.

30 Wittkampf KA, Naeije L, Schene A, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen Hosp Psychiatry* 2007; **29**: 388–95.

31 Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ* 2012; **184**: 191–6.

32 Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *Gen Hosp Psychiatry* 2015; **37**: 67–75.

33 Löwe B, Spitzer RL, Gräfe K, Kroenke K, Quenter A, Zipfel S, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disord* 2004; **78**: 131–40.

34 Lowe B, Kroenke K, Kerstin G. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *J Psychosom Res* 2005; **58**: 163–71.

35 Cholera R, Gaynes BN, Pence BW, Bassett J, Qangule N, Macphail C, et al. Validity of the Patient Health Questionnaire-9 to screen for depression in a high-HIV burden primary healthcare clinic in Johannesburg, South Africa. *J Affect Disord* 2014; **167**: 160–6.

36 Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; **155**: 529–36.

37 Mitchell AJ. Sensitivity × PPV is a recognized test called the clinical utility index (CUI +). *Eur J Epidemiol* 2011; **26**: 251–2.

38 Mitchell AJ. The clinical significance of subjective memory complaints in the diagnosis of mild cognitive impairment and dementia: a meta-analysis. *Int J Geriatr Psychiatry* 2008; **23**: 1191–202.

39 Mitchell AJ. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *J Psychiatr Res* 2009; **43**: 411–31.

40 Reeve JL, Lloyd-Williams M, Dowrick C. Revisiting depression in palliative care settings: the need to focus on clinical utility over validity. *Palliat Med* 2008; **22**: 383–91.

41 Li J, Fine JP. Assessing the dependence of sensitivity and specificity on prevalence in meta-analysis. *Biostatistics* 2011; **12**: 710–22.

42 Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988; **240**: 1285–93.

43 Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327**: 557–60.

44 Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Stat Med* 2006; **25**: 3443–57.

45 Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002; **21**: 1559–73.

46 Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med* 2010; **8**: 348–53.

47 Ayalon L, Goldfracht M, Bech P. 'Do you think you suffer from depression?' Reevaluating the use of a single item question for the screening of depression in older primary care patients. *Int J Geriatr Psychiatry* 2010; **25**: 497–502.

48 Azah N., Shah M, Juwita S, et al. Validation of the Malay version brief Patient Health Questionnaire (PHQ-9) among adult attending family medicine clinics. *Int Med J* 2005; **12**: 259–64.

49 Cannon DS, Tiffany ST, Coon H, Scholand MB, McMahon WM, Leppert MF. The PHQ-9 as a brief assessment of lifetime major depression. *Psychol Assess* 2007; **19**: 247–51.

50 Chen S, Fang Y, Chiu H, Fan H, Jin T, Conwell Y. Validation of the nine-item Patient Health Questionnaire to screen for major depression in a Chinese primary care population. *Asia Pac Psychiatry* 2013; **5**: 61–8.

51 Chen S, Chiu H, Xu B, Ma Y, Jin T, Wu M, et al. Reliability and validity of the PHQ-9 for screening late-life depression in Chinese primary care. *Int J Geriatr Psychiatry* 2010; **25**: 1127–33.

52 de Lima Osório F, Vilela Mendes A, Crippa JA, Loureiro SR. Study of the discriminative validity of the PHQ-9 and PHQ-2 in a sample of Brazilian women in the context of primary health care. *Perspect Psychiatr Care* 2009; **45**: 216–27.

53 Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, et al. Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. *Psychiatry Res* 2013; **210**: 653–61.

54 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**: 835–6.

55 Henkel V, Mergl R, Kohnen R, Allgaier AK, Moller HJ, Hegerl U. Use of brief depression screening tools in primary care: consideration of heterogeneity in performance in different patient groups. *Gen Hosp Psychiatry* 2004; **26**: 190–8.

56 Lamers F, 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**: 679–87.

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press

57 Liu SI, Yeh ZT, Huang HC, Sun FJ, Tjung JJ, Hwang LC, et al. Validation of Patient Health Questionnaire for depression screening among primary care patients in Taiwan. *Compr Psychiatry* 2011; **52**: 96–101.

58 Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry* 2008; **8**: 46.

59 Patel V, Araya R, Chowdhary N, King M, Kirkwood B, Nayak S, et al. Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires. *Psychol Med* 2008; **38**: 221–8.

60 Phelan E, Williams B, Meeker K, Bonn K, Frederick J, Logerfo J, et al. A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam Pract* 2010; **11**: 63.

61 Richardson LP, Rockhill C, Russo JE, Grossman DC, Richards J, McCarty C, et al. Evaluation of the PHQ-2 as a brief screen for detecting major depression among adolescents. *Pediatrics* 2010; **125**: e1097–103.

62 Sherina MS, Arroll B, Goodyear-Smith F. Criterion validity of the PHQ-9 (Malay version) in a primary care clinic in Malaysia. *Med J Malaysia* 2012; **67**: 309–15.

63 Spitzer RL, Kroenke K, Williams JBW. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA* 1999; **282**: 1737–44.

64 Sung SC, Low CC, Fung DS, Chan YH. Screening for major and minor depression in a multiethnic sample of Asian primary care patients: a comparison of the nine-item Patient Health Questionnaire (PHQ-9) and the 16-item Quick Inventory of Depressive Symptomatology – Self-Report (QIDS-SR16). *Asia Pac Psychiatry* 2013; **5**: 249–58.

65 Whooley MA, Avins AL, Miranda J, Browner WS. Case-finding instruments for depression: two questions are as good as many. *J Gen Intern Med* 1997; **12**: 439–45.

66 Wittkampf K, van Ravesteijn H, Baas K, van de Hoogen H, Schene A, Bindels P, et al. 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**: 451–9.

67 Yeung A, Fung F, Yu SC, Vorono S, Ly M, Wu S, et al. Validation of the Patient Health Questionnaire-9 for depression screening among Chinese Americans. *Compr Psychiatry* 2008; **49**: 211–7.

68 Zuithoff NP, Vergouwe Y, King M, Nazareth I, van Wezep MJ, Moons KG, et al. The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a crosssectional study. *BMC Fam Pract* 2010; **11**: 98.

69 Maxwell M, Harris F, Hibberd C, Donaghy E, Pratt R, Williams C, et al. A qualitative study of primary care professionals' views of case finding for depression in patients with diabetes or coronary heart disease in the UK. *BMC Fam Pract* 2013; **14**: 46.

https://doi.org/10.1192/bjpo.bp.115.001685 Published online by Cambridge University Press