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.

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.

Copyright and usage
© The Royal College of Psychiatrists 2016. This is an open access article distributed under the terms of the Creative Commons Non-Commercial, No Derivatives (CC BY-NC-ND) licence.

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.

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 individually many GPs have considerable experience in managing depression. 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 visit1 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 al5 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 depression11 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 application of a test to confirm those with a condition.15 Screening and case finding have been proposed as solutions adopted into 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).
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 advice24 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 CTPFPHC 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.33 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)34 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.35 This recommendation has recently been revised and extended.36 In the UK, the national guidelines have reversed their advice37 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.38 GPs in the UK have been less enthusiastic than patients about routine use of depression scales,39 leading to the removal of depression screening incentives from the UK QoF. In 2013, the Canadian CTPFPHC reconsidered its earlier guideline and also recommended against screening adults for depression in primary care settings.40 Thus, although some still advocate screening for depression, others do not and the argument has become polarised.41 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.42

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.33 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

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 × 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
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. 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 statistic which describes the percentage of total variation across studies that is caused by heterogeneity rather than chance. As per convention, we considered an 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. 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. 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. Details regarding the search results, including reasons for exclusion of articles are summarised in Fig. 1.

Fig. 1 PRISMA flow diagram of search strategy.
### 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-algorithm 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 al32 and Lowe et al33 compared all three diagnostic methods. Chen et al, Kroenke et al, Liu et al, de Lima Osorio 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 Zithoff 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 reference standard, with a mean patient age of 49.38 years, and 61% were female.

### 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.15,45,55,59 Three studies had either high risk of bias or were considered possible outliers. Richardson et al utilised adolescents seen in primary care; Whooley et al used the Whooley questions and was eventually excluded; finally Cannon et al 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

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

#### Additional information

**Table 1 Summary of included studies**

<table>
<thead>
<tr>
<th>Author</th>
<th>PHQ method</th>
<th>Sample mean age and % male/female</th>
<th>Sample size</th>
<th>Prevalence of depression, %</th>
<th>Reference standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowe et al35</td>
<td>PHQ-2 (linear) ≥2</td>
<td>42 years, 32.5% male</td>
<td>520</td>
<td>13.7</td>
<td>SCID, DSM-IV</td>
</tr>
<tr>
<td>Lowe et al31</td>
<td>PHQ-9-linear and algorithm</td>
<td>41.7 years, 32.9% male</td>
<td>501</td>
<td>13.2</td>
<td>SCID, DSM-IV</td>
</tr>
<tr>
<td>Arroll et al35</td>
<td>PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10, PHQ-9 (algorithm)</td>
<td>49 years, 39% male</td>
<td>2542</td>
<td>6.2</td>
<td>SCID, DSM-IV</td>
</tr>
<tr>
<td>Ayalon et al47</td>
<td>PHQ-9 (algorithm)</td>
<td>75 years, 59.5% male</td>
<td>153</td>
<td>3.9</td>
<td>SCID, DSM-IV</td>
</tr>
<tr>
<td>Azah et al48</td>
<td>PHQ-9 (linear) ≥10</td>
<td>38.7 years, 38.3% male</td>
<td>180</td>
<td>46.1</td>
<td>CIDI, ICD-10</td>
</tr>
<tr>
<td>Cannon et al39</td>
<td>PHQ-9 (linear)</td>
<td>57.2 years, 54% male</td>
<td>526</td>
<td>26.6</td>
<td>SCID DSM-IV MDD lifetime</td>
</tr>
<tr>
<td>Chen et al40</td>
<td>PHQ-9 (linear) ≥10</td>
<td>Age not reported, 47% male</td>
<td>262</td>
<td>16.7</td>
<td>SCID, DSM-IV</td>
</tr>
<tr>
<td>Chen et al41</td>
<td>PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10</td>
<td>68.5 years, 56.3% female</td>
<td>77</td>
<td>54.5</td>
<td>SCID DSM-IV</td>
</tr>
<tr>
<td>De Lima Osorio et al52</td>
<td>PHQ-9 (linear) ≥10; PHQ-2 (linear) ≥2</td>
<td>48% under 30, 52% between 31 and 50 years, 100% female</td>
<td>177</td>
<td>33.9</td>
<td>SCID DSM-IV</td>
</tr>
<tr>
<td>Gelaye et al53</td>
<td>PHQ-9 (linear) ≥10</td>
<td>35.1 years, 61.3% female</td>
<td>363</td>
<td>12.7</td>
<td>SCAN DSM-IV</td>
</tr>
<tr>
<td>Gilbody et al54</td>
<td>PHQ-9 (linear) ≥10</td>
<td>42.5 years, 77.1% female</td>
<td>96</td>
<td>37.5</td>
<td>SCID DSM-IV</td>
</tr>
<tr>
<td>Henkel et al36</td>
<td>PHQ-2 (linear) ≥2</td>
<td>53.9 years, 75% female</td>
<td>382</td>
<td>10.0</td>
<td>SCID DSM</td>
</tr>
<tr>
<td>Kroenke et al37</td>
<td>PHQ-9 (linear) ≥10</td>
<td>46 years, 66% female</td>
<td>580</td>
<td>7.1</td>
<td>DSMIV</td>
</tr>
<tr>
<td>Lamers et al49</td>
<td>PHQ-9 (algorithm), PHQ-9 (linear) ≥8</td>
<td>71.4 years, 51.8% male</td>
<td>713</td>
<td>17.3</td>
<td>MINI DSM-IV</td>
</tr>
<tr>
<td>Liu et al50</td>
<td>PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10</td>
<td>18 years or older, 39% male</td>
<td>1532</td>
<td>3.3</td>
<td>SCAN DSM-IV</td>
</tr>
<tr>
<td>Lotrakul et al51</td>
<td>PHQ-9 (linear) ≥10, PHQ-9 (algorithm)</td>
<td>45 years, 73.7% female</td>
<td>279</td>
<td>6.8</td>
<td>MINI DSM-IV</td>
</tr>
<tr>
<td>Patel et al52</td>
<td>PHQ-9 (linear) ≥10</td>
<td>37.5 years, 56.4% female</td>
<td>598</td>
<td>5.5</td>
<td>ICD-10</td>
</tr>
<tr>
<td>Phelan et al53</td>
<td>PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10</td>
<td>78 years, 62% female</td>
<td>69</td>
<td>11.6</td>
<td>SCID DSM</td>
</tr>
<tr>
<td>Richardson et al54</td>
<td>PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10</td>
<td>15.3 years, 60% female</td>
<td>442</td>
<td>4.3</td>
<td>DIS for MDD in children (DISC)</td>
</tr>
<tr>
<td>Sherma et al55</td>
<td>PHQ-9 (linear) ≥10</td>
<td>30.9 years, 100% female</td>
<td>146</td>
<td>12.1</td>
<td>CIDI, ICD-10</td>
</tr>
<tr>
<td>Spitzer et al56</td>
<td>PHQ-9 (algorithm)</td>
<td>46 years, 66% female</td>
<td>585</td>
<td>10.0</td>
<td>DSM</td>
</tr>
<tr>
<td>Sung et al57</td>
<td>PHQ-9 (linear) ≥10</td>
<td>36.1 years, 65.3% female</td>
<td>400</td>
<td>3.0</td>
<td>MINI DSM</td>
</tr>
<tr>
<td>Wittkampf et al58</td>
<td>PHQ-9 (linear) ≥10; PHQ-9 (algorithm)</td>
<td>49.8 years, 67.6% female</td>
<td>664</td>
<td>12.3</td>
<td>SCID-I</td>
</tr>
<tr>
<td>Yeung et al59</td>
<td>PHQ-9 (linear) ≥15</td>
<td>Not reported</td>
<td>184</td>
<td>22.8</td>
<td>SCID DSM</td>
</tr>
<tr>
<td>Zithoff et al60</td>
<td>PHQ-2 (linear) ≥2; PHQ-9 (linear) ≥10, PHQ-9 (algorithm)</td>
<td>51 years, 37% female</td>
<td>1352</td>
<td>13.0</td>
<td>CIDI DSM-IV</td>
</tr>
</tbody>
</table>

CIDI, Composite International Diagnostic interview; DIS, Diagnostic Interview Schedule; DSC, Diagnostic Interview Schedule for Children; DSM, Diagnostic and Statistical Manual of Mental Disorders; ICD, International Classification of Diseases; MIN, Mini International Neuropsychiatric Interview; PHQ, Patient Health Questionnaire; SCAN, Schedules for Clinical Assessment in Neuropsychiatry; SCID, Structured Clinical Interview for DSM Disorders.
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 sub-analysis 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.
### Table 3  Summary of Patient Health Questionnaire (PHQ) analysis in primary care

<table>
<thead>
<tr>
<th>Test</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>PPV</th>
<th>NPV</th>
<th>ROC</th>
<th>CUI+</th>
<th>CUI−</th>
<th>LR+</th>
<th>LR−</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main results</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-9-linear n=20</td>
<td>82.2 (74.3–88.9)</td>
<td>84.7 (80.8–89.8)</td>
<td>38.0% (36.1–39.9%)</td>
<td>97.7% (97.3–98.0%)</td>
<td>0.910 (0.892–0.930)</td>
<td>0.312 (0.311–0.313)</td>
<td>0.827 (0.826–0.828)</td>
<td>5.37 (5.09–5.67)</td>
<td>0.21 (0.19–0.24)</td>
</tr>
<tr>
<td>PHQ-9-algorithm n=9</td>
<td>58.4 (44.5–71.7)</td>
<td>92.1 (85.9–96.6)</td>
<td>47.4% (44.7–50.1%)</td>
<td>94.8% (94.3–95.3%)</td>
<td>0.733 (0.676–0.795)</td>
<td>0.277 (0.276–0.278)</td>
<td>0.873 (0.873–0.873)</td>
<td>7.39 (6.77–8.07)</td>
<td>0.45 (0.42–0.49)</td>
</tr>
<tr>
<td>PHQ-2 n=11</td>
<td>89.9 (83.4–94.9)</td>
<td>72.6 (66.0–78.7)</td>
<td>23.1% (21.5–24.7%)</td>
<td>97.5% (97.0–97.9%)</td>
<td>0.860 (0.819–0.903)</td>
<td>0.188 (0.187–0.189)</td>
<td>0.708 (0.707–0.717)</td>
<td>‘very poor’</td>
<td>‘excellent’</td>
</tr>
<tr>
<td><strong>Head-to-head results</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-9-linear n=8</td>
<td>87.0 (75.8–95.1)</td>
<td>87.2 (81.1–92.2)</td>
<td>38.2% (35.1–41.2%)</td>
<td>96.6% (98.3–98.9%)</td>
<td>0.920 (0.915–0.924)</td>
<td>0.322 (0.321–0.323)</td>
<td>0.877 (0.876–0.878)</td>
<td>7.66 (7.04–8.33)</td>
<td>0.18 (0.14–0.22)</td>
</tr>
<tr>
<td>PHQ-9-linear n=8</td>
<td>91.5 (83.6–96.9)</td>
<td>72.2 (64.0–79.8)</td>
<td>22.7% (20.8–24.6%)</td>
<td>99.0% (98.7–99.3%)</td>
<td>0.900 (0.865–0.934)</td>
<td>0.205 (0.205–0.206)</td>
<td>0.742 (0.741–0.743)</td>
<td>‘very poor’</td>
<td>‘excellent’</td>
</tr>
<tr>
<td>PHQ-9-algo n=6</td>
<td>53.0 (36.4–69.3)</td>
<td>95.7 (93.5–97.5)</td>
<td>58.4% (54.1–62.8%)</td>
<td>94.1% (93.5–94.7%)</td>
<td>0.715 (0.628–0.815)</td>
<td>0.273 (0.272–0.275)</td>
<td>0.905 (0.904–0.906)</td>
<td>12.35 (10.56–14.43)</td>
<td>0.55 (0.51–0.56)</td>
</tr>
<tr>
<td><strong>Moderator analysis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-9-linear n=16 Higher quality; same cut-off; adults</td>
<td>81.3 (71.6–89.3)</td>
<td>85.3 (81.0–89.1)</td>
<td>44.2% (41.9–46.6%)</td>
<td>38.9 (36.8–41.0)</td>
<td>97.5 (97.2–97.9)</td>
<td>0.316 (0.315–0.317)</td>
<td>0.832 (0.832–0.832)</td>
<td>5.53 (5.21–5.87)</td>
<td>0.21 (0.19–0.25)</td>
</tr>
<tr>
<td>PHQ-9-algorithm n=8 Higher quality; same cut-off; adults</td>
<td>56.8 (41.2–71.8)</td>
<td>93.3 (87.5–97.3)</td>
<td>60.3% (57.0–63.6%)</td>
<td>48.3% (45.4–51.3)</td>
<td>95.1% (94.7–95.6)</td>
<td>0.275 (0.274–0.276)</td>
<td>0.887 (0.887–0.887)</td>
<td>8.46 (9.07–8.46)</td>
<td>0.46 (0.43–0.49)</td>
</tr>
<tr>
<td><strong>PHQ-2 n=9 Higher quality; same cut-off; adults</strong></td>
<td>89.3 (81.5–95.1)</td>
<td>75.9% (70.1–81.3)</td>
<td>27.7% (25.8–29.6%)</td>
<td>25.5% (24.3–28.3%)</td>
<td>98.6% (98.3–99.0%)</td>
<td>0.236 (0.235–0.237)</td>
<td>0.749 (0.749–0.749)</td>
<td>3.71 (3.52–3.90)</td>
<td>0.14 (0.11–0.18)</td>
</tr>
</tbody>
</table>

PPV, positive predictive value; NPV, negative predictive value; ROC, receiver operating characteristic; CUI+, positive clinical utility index; CUI−, negative clinical utility index; LR+, positive likelihood ratio; LR−, negative likelihood ratio.

*a. Alternative calculation based on bivariate calculation in STATA.*
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 ≥2 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
**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-algorithm v. PHQ-2.
<table>
<thead>
<tr>
<th>Test</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>PPV</th>
<th>NPV</th>
<th>CUI+</th>
<th>CUI-</th>
<th>LR+</th>
<th>LR−</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHQ-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cut PHQ-2 ≥ 1 n=20</td>
<td>96.05 (92.29–98.60)</td>
<td>52.18 (43.42–60.8)</td>
<td>14.7 (13.6–15.9)</td>
<td>99.4 (99.1–99.6)</td>
<td>0.141 (0.140–0.142)</td>
<td>‘very poor’</td>
<td>0.519 (0.519–0.519)</td>
<td>2.01 (1.95–2.07)</td>
</tr>
<tr>
<td>Cut PHQ-2 ≥ 2 n=9</td>
<td>92.20 (85.21–97.10)</td>
<td>70.98 (64.63–76.94)</td>
<td>22.8 (21.2–24.5)</td>
<td>99.0 (98.7–99.3)</td>
<td>0.211 (0.210–0.212)</td>
<td>‘very poor’</td>
<td>0.703 (0.703–0.703)</td>
<td>3.18 (3.04–3.32)</td>
</tr>
<tr>
<td>Cut PHQ-2 ≥ 3 n=11</td>
<td>76.22 (61.1–88.53)</td>
<td>88.66 (85.01–91.86)</td>
<td>38.6 (35.9–41.3)</td>
<td>97.6 (97.2–97.9)</td>
<td>0.294 (0.293–0.295)</td>
<td>‘very poor’</td>
<td>0.865 (0.865–0.865)</td>
<td>6.74 (6.23–7.30)</td>
</tr>
<tr>
<td>Cut PHQ-2 ≥ 4 n=11</td>
<td>61.46 (44.00–77.52)</td>
<td>94.14 (91.73–96.15)</td>
<td>53.4 (49.7–57.2)</td>
<td>95.7 (95.2–96.3)</td>
<td>0.329 (0.328–0.330)</td>
<td>‘very poor’</td>
<td>0.901 (0.901–0.901)</td>
<td>10.45 (9.22–11.85)</td>
</tr>
<tr>
<td>Cut PHQ-2 ≥ 5 n=11</td>
<td>47.33 (26.78–68.37)</td>
<td>97.60 (95.36–99.12)</td>
<td>72.8 (67.2–78.4)</td>
<td>93.2 (92.2–94.1)</td>
<td>0.344 (0.241–0.436)</td>
<td>‘very poor’</td>
<td>0.909 (0.908–0.910)</td>
<td>19.77 (15.23–25.68)</td>
</tr>
<tr>
<td>Cut PHQ-2 ≥ 6 n=11</td>
<td>51.80 (23.07–79.88)</td>
<td>98.63 (96.91–99.66)</td>
<td>83.3 (78.5–88.2)</td>
<td>93.5 (92.6–94.4)</td>
<td>0.421 (0.419–0.424)</td>
<td>‘very poor’</td>
<td>0.421 (0.419–0.424)</td>
<td>0.922 (0.921–0.923)</td>
</tr>
</tbody>
</table>

| 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 (0.258–0.259) | ‘very poor’ | 0.611 (0.611–0.611) | 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) | 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 (0.741–0.743) | 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) | 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) | 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) | 6.23 (5.74–6.76) | 0.28 (0.25–0.32) |
| Cut PHQ-9 ≥ 12 n=20 | 68.37 (54.71–80.38) | 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 (0.870–0.870) | 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) | ‘very poor’ | 0.421 (0.419–0.423) | 9.84 (8.38–11.55) | 0.32 (0.28–0.36) |
| 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) | ‘very poor’ | 0.898 (0.897–0.898) | 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; LR+, positive likelihood ratio; LR−, 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).
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 PHQ-9-algorithm 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, 93.3% are true negatives (true non-cases). Similarly, 96.3% 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.¹⁸²⁶ 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.

**Acknowledgements**

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

**References**


Mitchell AJ. Sensitivity × PPV is a recognized test called the clinical utility index (CUI +). Eur J Epidemiol 2011; 26: 251–1.


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: f7143.


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.


