Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis

Laura Manea MSc, Simon Gilbody PhD, Dean McMillan PhD

See related commentary by Kroenke at www.cmaj.ca/lookup/doi/10.1503/cmaj.112004

Abstract

Background: The brief Patient Health Questionnaire (PHQ-9) is commonly used to screen for depression with 10 often recommended as the cut-off score. We summarized the psychometric properties of the PHQ-9 across a range of studies and cut-off scores to select the optimal cut-off for detecting depression.

Methods: We searched Embase, MEDLINE and PsycINFO from 1999 to August 2010 for studies that reported the diagnostic accuracy of PHQ-9 to diagnose major depressive disorders. We calculated summary sensitivity, specificity, likelihood ratios and diagnostic odds ratios for detecting major depressive disorder at different cut-off scores and in different settings. We used random-effects bivariate meta-analysis at cut-off points between 7 and 15 to produce summary receiver operating characteristic curves.

Results: We identified 18 validation studies (n = 7180) conducted in various clinical settings. Eleven studies provided details about the diagnostic properties of the questionnaire at more than one cut-off score (including 10), four studies reported a cut-off score of 10, and three studies reported cut-off scores other than 10. The pooled specificity results ranged from 0.73 (95% confidence interval [CI] 0.63–0.82) for a cut-off score of 7 to 0.96 (95% CI 0.94–0.97) for a cut-off score of 15. There was major variability in sensitivity for cut-off scores between 7 and 15. There were no substantial differences in the pooled sensitivity and specificity for a range of cut-off scores (8–11).

Interpretation: The PHQ-9 was found to have acceptable diagnostic properties for detecting major depressive disorder for cut-off scores between 8 and 11. Authors of future validation studies should consistently report the outcomes for different cut-off scores.

Competing interests: Dean McMillan has received grants from the UK National Institute for Health Research. No competing interests declared by Laura Manea and Simon Gilbody.

This article has been peer reviewed.

Correspondence to:
Laura Manea,
lem514@york.ac.uk

***CMAJ* 2012. DOI:10.1503/cmaj.110829**

Depressive disorders are still underrecognized in medical settings despite major associated disability and costs. The use of short screening questionnaires may improve the recognition of depression in different medical settings.[1] The depression module of the Patient Health Questionnaire (PHQ-9) has become increasingly popular in research and practice over the past decade.[2] In its initial validation study, a score of 10 or higher had a sensitivity of 88% and a specificity of 88% for detecting major depressive disorders. Thus, a score of 10 has been recommended as the cut-off score for diagnosing this condition.[3]

In a recent review of the PHQ-9, Kroenke and colleagues argued against inflexible adherence to a single cut-off score.[2] A recent analysis of the management of depression in general practice in the United Kingdom showed that the accuracy of predicting major depressive disorder could be improved by using 12 as the cut-off score.[4]

Given the widespread use of PHQ-9 in screening for depression and that certain cut-off scores are being recommended as part of national strategies to screen for depression (based on initial validation studies, which might not be generalizable),[4,5] we attempted to determine whether the cut-off of 10 is optimum for screening for depression. This question could not be answered by two previous systematic reviews[6,7] because of the small number of primary studies available at the time. We also aimed to provide greater clarity about the proper use of PHQ-9 given the many settings in which it is used.

Methods

We performed a meta-analysis of the available literature using recently developed bivariate meta-analysis methods.[8,9,10] We included all cross-sectional validation studies of PHQ-9 as a screening tool for major depressive disorder that met our inclusion criteria.

Literature search

We searched Embase, MEDLINE and PsycINFO from 1999 (when PHQ-9 was issued) to August 2010 using the terms "PHQ-9" and "patient health questionnaire." We manually searched the reference lists of each study that met our inclusion criteria, and we performed a reverse citation search in Web of Science to identify additional studies. We contacted the authors of original studies to obtain unpublished data if necessary. We also contacted the authors of unpublished studies and conference abstracts, and we reviewed these in an attempt to minimize publication bias. No publication or language restrictions were applied.

Study selection

We selected studies that reported the accuracy of PHQ-9 for diagnosing major depressive disorder. The studies had to provide sufficient data to allow us to calculate contingency tables. We included studies that defined major depressive disorder according to standard classification systems such as the International Classification of Diseases (ICD) or the Diagnostic and Statistical Manual of Mental Disorders (DSM). We excluded studies in which the diagnoses were not made using a standardized diagnostic interview schedule (e.g., 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]). We made the final selection after examining the full-text articles. Figure 1 shows selection of studies using this strategy.

Data abstraction

We collected information about study characteristics and quality using a standardized data collection form. We included the following characteristics: settings, year of study, sample size, study design, timing between reference and index tests, training of the rater of the reference test, blinding of the assessor of the reference test, data integrity, cut-off score, and translation and validation of non-English versions of PHQ-9. We recorded accuracy data for the reported cut-off scores in contingency tables.

Quality assessment

We based our assessment of quality on current guidelines for evaluating diagnostic studies.[11] We followed the quality ratings used in other studies of the diagnostic characteristics of psychological measures to generate our quality assessment criteria.[12]

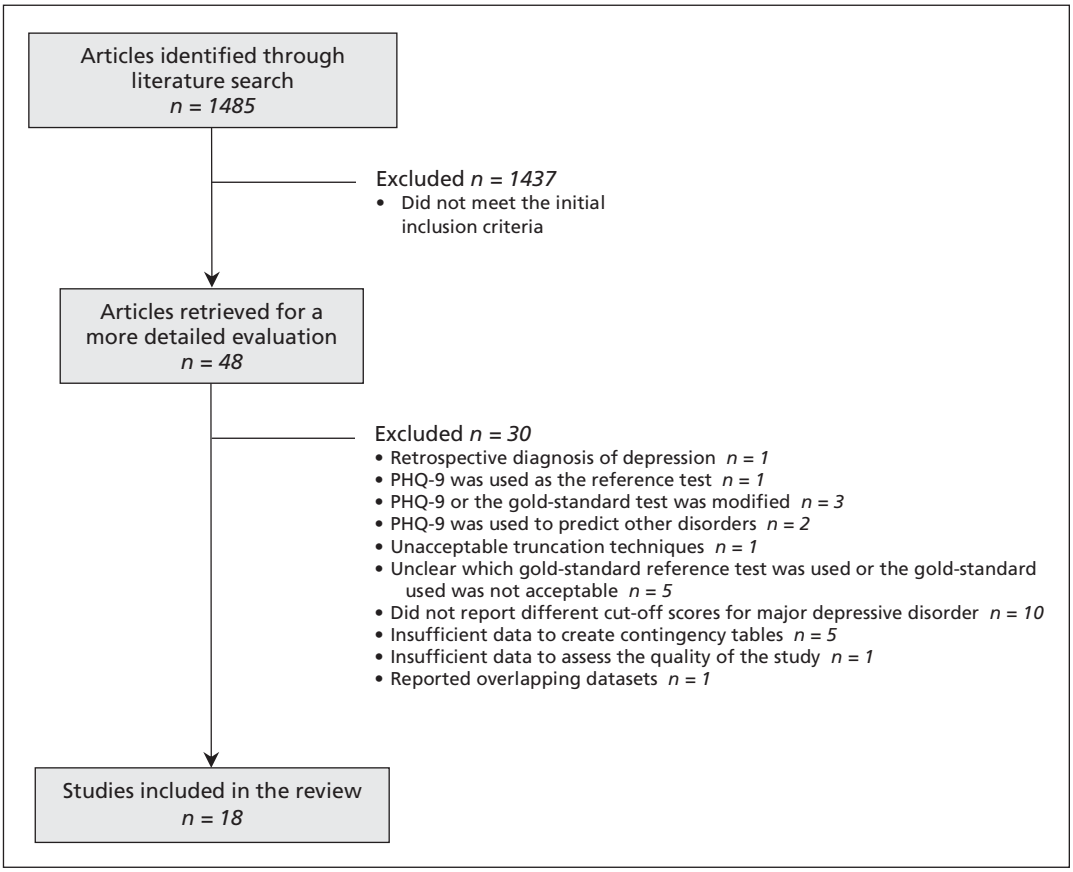

Figure 1: Selection of studies for inclusion in the systematic review. Note: PHQ-9 = Patient Health Questionnaire.

We devised specific questions to evaluate the translation and validation of non-English versions of PHQ-9 and design issues that could influence the outcomes.

Our quality criteria included adequate sample size ($n \geq 250$). We considered the single-gate design to be ideal. A single-gate study compares the results from an index test with those from a reference standard used to confirm the diagnosis. In a two-gate design, the disease status is already known. A two-gate study compares the results of an index test for patients with an established diagnosis of the target condition, which are therefore treated as the reference standard, with the results of the same test in healthy people or people with another condition.[13]

We assessed whether the rater of the test had been trained in the use of the reference test and whether the assessor was blinded to the result of the reference test. We also examined whether withdrawals or drop-outs were explained or accounted for. We considered a rate of less than 20% refusals or drop-outs to be acceptable. It was also important that the time between the index test and the reference test to be two weeks or less. We considered the chosen cut-off score for reporting to be acceptable if it was tested as best trade-off. Finally, for studies that used a translated version of PHQ-9, we examined whether the translation had been validated according to recognized standards.[14,15,16]

Data synthesis and statistical analysis

We constructed 2 × 2 tables for each cut-off score and computed the sensitivity, specificity and positive and negative predictive values.

We performed a bivariate meta-analysis to obtain pooled estimates of specificity and sensitivity and their associated 95% confidence intervals (CIs).[17] We constructed summary receiver operating characteristic curves[17] using the bivariate model to produce a 95% confidence ellipse within the receiver operating characteristic curve space.[18] Each data score in this space represents a separate study. This is unlike a traditional receiver operating characteristic plot, which explores the effect of varying thresholds on sensitivity and specificity in a single study.

We assessed between-study heterogeneity using the I^2 statistic for the pooled diagnostic odds ratio (OR),[19] which describes the percentage of total variation across studies that is caused by heterogeneity rather than chance. 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. We identified the studies that were outside of the 95% confidence ellipse by visually inspecting the summary receiver operating characteristic curve plots.

We performed a meta-regression analysis of the logit diagnostic ORs using a priori identified sources of heterogeneity entered as covariates in the meta-regression model.[9] We investigated heterogeneity resulting from the characteristics of the sample or study design by exploring the effects of potential predictive variables.[8]

We examined publication and small study bias using Begg's funnel plots of log diagnostic ORs versus the inverse of the variance.[10,20]

Results

We found 18 studies (7180 participants) that met our inclusion criteria (Appendix 1, www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.110829/-/DC1). The excluded studies and the reasons for exclusion are described in Appendix 2 (available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.110829/-/DC1). Eight of the included studies validated PHQ-9 in primary care;[21–28] five in specialized secondary care services (brain injury,[29] cardiology,[30,31] stroke[32] and renal[33]) and two in samples from the community.[34,35] Three studies were conducted in mixed settings (outpatient clinics and family practices).[3,36,37] The mean age of participants ranged from 24.8 to 71.4 years.[23,34] Within the 18 included studies, the prevalence of depression, as diagnosed by the gold-standard tests, ranged from 2.5% to 37.5%.[22,34] The included studies used the English version[3,22,28–35] and translated versions (Portuguese,[21] German,[36,37] Dutch,[23,27] Thai,[24] Malay[25] and Konkani[26]) of PHQ-9.

Quality assessment

All included studies used the DSM or ICD-10 diagnosis of depression, established using a standardized interview schedule. The interview schedules used were the English or translated versions of the MINI,[23,24,30,34] SCID,[3,21,22,27–29,32,33,35–37] CIDI,[25] DIS[31] or CIS-R.[26] In 14 studies, participants were assessed by use of both the PHQ-9 and reference test.[3,21,22,24–26,29,35–37] Four studies used other study designs in which only patients who scored below a certain cut-off on the PHQ-9[23,27,28] or who showed the core symptoms or at least two symptoms on the PHQ-9[32] underwent testing using the gold-standard test.

Meta-analysis

Eighteen studies (7180 patients, 927 with major depressive disorder confirmed by DSM or ICD) reported the diagnostic properties of PHQ-9 at different cut-off scores. We pooled studies with different cut-off scores because not all studies reported the same cut-off scores.

We found a high level of between-study heterogeneity for psychometric attributes (I^2 = 82.4%). The pooled sensitivity for a cut-off score of 10 was 0.85 (95% CI 0.75–0.91) and the pooled specificity was 0.89 (95% CI 0.83–0.92) (Table 1).

When we summarized individual studies within receiver operating characteristic curve space for a cut-off score of 10, we found that most studies gathered within an informative top left corner (Appendix 3, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.110829/-/DC1).

Two of the three studies with a relatively low sensitivity at a cut-off score of 10 were conducted in cardiology settings.[30,31] The other study was conducted in primary care and was the only one that used a diagnostic interview based on ICD criteria.[25] The pooled sensitivity for studies performed in hospital settings (0.74, 95% CI 0.55–0.86) was lower than that for primary care (0.89, 95% CI 0.66–0.97); however, the specificity was very similar for these settings (0.89 [95% CI 0.87–0.91] v. 0.88 [95% CI 0.80–0.93])

Because of the between-study heterogeneity, we performed a meta-regression. The criterion of blind application of a diagnostic gold standard was the only a priori source of heterogeneity that was predictive (p = 0.032). Diagnostic performance did not vary according to the percentage of women (p = 0.39), study setting (primary care and community settings v. hospital; p = 0.73), prevalence of depression (p = 0.70), sample size (p = 1.00) or mean age (p = 0.28).

Translation of PHQ-9 (p = 0.33), study design (single gated v. double gated; p = 0.19), timing of index and reference testing (p = 0.61), standard of data integrity reporting (p = 0.40) and training of the rater of the reference test (p = 0.11) did not have a significant impact on diagnostic performance.

We found that the diagnostic OR was lower in hospital settings[30–33] (diagnostic OR 25.43, 95% CI 11.35–57.00) than in primary care settings[21,22,25–27] (diagnostic OR 65.26, 95% CI 9.17–464.47) (Table 2). Studies in primary care and hospital settings were equally heterogeneous (primary care I^2 = 84.7%; hospital I^2 = 84.2%).

Alternative cut-off scores

Fourteen studies reported the diagnostic properties of PHQ-9 for a cut-off score of 10, and some reported additional cut-off scores (Table 1). A cut-off score greater than 10 was reported as being optimal by only three studies.[3,25,34] In 4 studies, a cut-off score of 11 or 12 had better diagnostic properties than a cut-off score of 10.[22,29,36,37]

The pooled sensitivity and specificity results show no significant differences in the diagnostic properties of PHQ-9 for cut-off scores between 8 and 11. The pooled specificity was between 0.73 (95% CI 0.63–0.82) for a cut-off score of 7 and 0.96 (95% CI 0.94–0.97) for a cut-off score of 15. Overall, the sensitivity did not decrease as the

Table 1: Pooled estimates of the sensitivity, specificity, positive and negative likelihood ratios and diagnostic odds ratios of the brief Patient Health Questionnaire (PHQ-9) for diagnosing major depressive disorder, by cut-off score

Cut-off score	No. of studies	No. of patients	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)
7	5	2794	0.83 (0.70–0.91)	0.73 (0.63–0.82)	3.20 (2.21–4.63)	0.22 (0.12–0.41)	14.18 (6.21–32. 33)
8	6	3306	0.82 (0.66–0.92)	0.83 (0.69–0.92)	5.17 (2.30–11.58)	0.20 (0.08–0.47)	25.46 (5.34–121.38)
9	7	3269	0.83 (0.68–0.92)	0.86 (0.76–0.92)	6.07 (3.35–11.01)	0.18 (0.08–0.39)	32.23 (10.06–103. 25)
10	16	5782	0.85 (0.75–0.91)	0.89 (0.83–0.92)	7.83 (5.22–11.74)	0.16 (0.09–0.27)	47.50 (22.94–98.35)
11	10	3451	0.89 (0.75–0.96)	0.89 (0.79–0.94)	8.43 (4.29–16.58)	0.11 (0.04–0.29)	75.03 (25.29–222.53)
12	10	3451	0.77 (0.60–0.88)	0.91 (0.84–0.95)	8.86 (5.65–13.90)	0.24 (0.13 - 0.44)	36.15 (20.16–64.85)
13	8	2759	0.81 (0.75–0.85)	0.92 (0.86–0.95)	10.19 (6.19 - 16.77)	0.20 (0.15–0.26)	50.03 (29.81–83.97)
14	6	2162	0.67 (0.57–0.76)	0.95 (0.90–0.97)	14.23 (7.10 - 28.52)	0.33 (0.24–0.45)	42.13 (18.73–94.74)
15	6	2482	0.62 (0.48–0.74)	0.96 (0.94–0.97)	18.57 (11.37-30.33)	0.39 (0.27–0.55)	47.60 (23.09–98.14)

Note: CI = confidence interval.

cut-off score increased (Table 2). There was a decrease in sensitivity to 0.77 (95% CI 0.60–0.88) for a cut-off score of 12; however, sensitivity improved to 0.81 for a cut-off score of 13 (95% CI 0.75–0.85). A cut-off score of 11 had the best trade-off between sensitivity and specificity.

Interpretation

We found that PHQ-9 has acceptable diagnostic properties at a range of cut-off scores (8–11). There were no significant differences in sensitivity or specificity at a cut-off score of 10 compared with other cut-off scores within this interval (8–11). However, our results are based on a variable number of studies for each cut-off scores. To provide a more valid comparison between different cut-off scores, more studies that consistently report data for a range of cut-off scores are needed.

This systematic review of the diagnostic properties of the PHQ-9 for different cut-off scores follows previous recommendations to analyze the sensitivity and specificity of the PHQ-9 at various cut-off scores using bivariate meta-analysis.[6,7] Our results support previous findings that PHQ-9 has acceptable diagnostic properties for major depressive disorder. Its diagnostic accuracy was reasonably consistent despite clinical heterogeneity of the included studies.

The methodologic quality of the studies varied, and the level of between-study heterogeneity was consistently high. A significant finding is that the reported blind application of a diagnostic gold standard was the only predictive source of heterogeneity. Although no other a priori sources of heterogeneity apart from blinding were able to explain the substantial between-study variation, we recommend that the proposed potential sources of heterogeneity (e.g., single-gated study design, training of the rater of the reference test, blinding of the assessor to the result of the reference test, and the use of validated translations of the index and reference tests) should be included if further primary studies are performed.

We found that there were no significant differences in pooled sensitivity and specificity for cut-off scores between 8 and 11. There was a decrease in sensitivity for a cut-off score of 12; however, sensitivity improved for a cut-off score of 13. The fact that different studies contributed to the calculations of different cut-off scores might be a possible explanation for this unexpected trend.

Limitations

Our study has several limitations. First, we could not rule out publication bias. Study selection was perfomed by one author, and this might have introduced bias. Our quality assessment criteria have not been validated. We were unable to fully explain the large amount of heterogeneity between studies, and caution should be used when interpreting the results.

Four of the studies included in our analysis used designs in which only patients who scored below a certain cut-off on the PHQ-9[23,27,28] or who showed the core symptoms or at least two symptoms on the PHQ-9[32] underwent testing using the gold-standard test. These designs may have led to partial verification bias. By selectively including patients with a known or possible diagnosis of depression, the prevalence of depressive disease in the sample could be increased and the sensitivity may have been overestimated.

Conclusions and implications for further research

The PHQ-9 is a popular tool for detecting depression in many settings. Our findings emphasize the importance of using caution when choosing a specific cut-off score, taking into account the characteristics of the population, the settings and the efficacy of screening on outcomes. A cut-off score of 10 may result in many false negatives in hospital settings, while more false-positive results may be seen in primary care. Our results support previous observations that a cut-off score of 10 is not superior to a score of 11 or 12 in terms of sensitivity.[4,22,37] However, the included studies reported different cut-off scores, and some of them made a post hoc selection of the best cut-off score. The

Table 2: Pooled estimates of the sensitivity, specificity, positive and negative likelihood ratios and diagnostic odds ratios of the brief Patient Health Questionnaire (PHQ-9) for diagnosing major depressive disorder, by setting

Setting	No. of studies	No. of patients	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)
Primary care	6	1994	0.89 (0.66–0.97)	0.88 (0.80–0.93)	7.56 (3.93–14.55)	0.11 (0.02–0.45)	65.26 (9.17–464.47)
Hospital	5	1730	0.74 (0.55–0.86)	0.89 (0.87–0.91)	7.29 (5.68–9.37)	0.28 (0.15–0.52)	25.43 (11.35–57.00)

Note: CI = confidence interval.

results of our analysis are based on a different number of studies for each cut-off score, each of which included a different population and had different methodologic characteristics. We recommend that studies of diagnostic accuracy should report the results for all cut-off scores and avoid reporting only the scores that are determined to be best after the study has been performed.

It has previously been suggested that the blinding of the assessor of the reference test could account for some of the between-study variation;[6] however, this has not been shown empirically. Blinding is one of the few evidence-based quality criteria that can have a significant effect on estimates of diagnostic accuracy.[38] Our study provides further evidence that future studies of diagnostic accuracy should explicitly report blinding.

It is also important that researchers recruit participants who represent the intended spectrum of severity of the target condition and that all participants complete both the index and reference test.

The same cut-off score might not be appropriate in all settings. PHQ-9 is a highly useful screening tool, but it is not a stand-alone diagnostic test. Given the structure of the questionnaire and its intended use as a screening tool, the optimal cut-off score may differ depending on the setting

References

1. Wright AF. Should general practitioners be testing for depression? *Br J Gen Pract* 1994;44:132-5.
2. Kroenke K, Spitzer RL, Williams JBW, et al. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *Gen Hosp Psychiatry* 2010;32:345-59.
3. Kroenke K, Spitzer R, Williams J. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606-13.
4. Kendrick T, Dowrick C, McBride A, et al. Management of depression in UK general practice in relation to scores on depression severity questionnaires: analysis of medical record data. *BMJ* 2009;338:b750.
5. Clark DM, Layard R, Smithies R, et al. Improving access to psychological therapy: initial evaluation of two UK demonstration sites. *Behav Res Ther* 2009;47:910-20.
6. Gilbody S, Richards D, Brealey S, et al. Screening for depression in medical settings with the patient health questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* 2007;22:1596-602.
7. Wittkampf KA, Naeije L, Schene A, et al. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen Hosp Psychiatry* 2007;29:388-95.
8. Lijmer JG, Bossuyt PMM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21:1525-37.
9. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21:1559-73.
10. Song F, Khan KS, Dinnes J, et al. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol* 2002;31:88-95.
11. Devillé WL, Buntinx F, Bouter LM, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol* 2002;2:9.
12. Mitchell AJ. Are one or two simple questions sufficient to detect depression in cancer and palliative care? A Bayesian meta-analysis. *Br J Cancer* 2008;98:1934-43.
13. Centre for Reviews and Dissemination. *Systematic reviews: CRD's guidance for undertaking reviews in health care.* York (UK): University of York; 2009.
14. Rahman A, Iqbal Z, Waheed W, et al. Translation and cultural adaptation of health questionnaires. *J Pak Med Assoc* 2003;53:142-7.
15. Brislin RW. Back-translation for cross-cultural research. *J Cross Cult Psychol* 1970;1:185-216.
16. Neewoor S, D'Uva F, Le Halpere A, et al. Assessing anxiety and depression on an international level [abstract PMC72]. *Value Health* 2009;12:A32-A33.
17. Reitsma JB, Glas AS, Rutjes AW, et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982-90.
18. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002; 21:1237-56.
19. Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.
20. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005; 58:882-93.
21. De Lima Osório F, Vilela Mendes A, Crippa J, et al. 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.
22. Gilbody S, Richards D, Barkham M. Diagnosing depression in primary care using self-completed instruments: UK validation of PHQ-9 and CORE-OM. *Br J Gen Pract* 2007;57:650-2.
23. Lamers F, Jonkers CCM, Bosma H, et al. 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.
24. Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry* 2008;8:46.
25. 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.
26. Patel V, Araya R, Chowdhary N, et al. Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires. *Psychol Med* 2008;38:221-8.
27. Wittkampf K, van Ravesteijn H, Baas K, 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.
28. Yeung A, Fung F, Yu S, et al. Validation of the Patient Health Questionnaire-9 for depression screening among Chinese Americans. *Compr Psychiatry* 2008;49:211-7.
29. Fann JR, Bombardier C, Dikmen S, et al. Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. *J Head Trauma Rehabil* 2005;20:501-11.
30. Stafford L, Berk M, Jackson H. Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients with coronary artery disease. *Gen Hosp Psychiatry* 2007;29:417-24.
31. Thombs BD, Ziegelstein RC, Whooley MA. Optimizing detection of major depression among patients with coronary artery disease using the Patient Health Questionnaire: data from the Heart and Soul Study. *J Gen Intern Med* 2008;23:2014-7.
32. Williams LS, Brizendine E, Plue L, et al. Performance of the PHQ-9 as a screening tool for depression after stroke. *Stroke* 2005;36:635-8.
33. Watnick S, Wang P, Demadura T, et al. Validation of 2 depression screening tools in dialysis patients. *Am J Kidney Dis* 2005; 46:919-24.
34. Adewuya AO, Ola BA, Afolabi OO. Validity of the Patient Health Questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *J Affect Disord* 2006;96:89-93.
35. Gjerdingen D, Crow S, McGovern P, et al. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *Ann Fam Med* 2009;7:63-70.
36. Gräfe K, Zipfel S, Herzog W, et al. Screening for psychiatric disorders with the Patient Health Questionnaire (PHQ). Results from the German validation study. *Diagnostica* 2004;50:171-81.
37. Löwe B, Spitzer R, Grafe K, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disord* 2004;78:131-40.
38. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999; 282:1061-6.

Affiliation: From the Department of Health Sciences, York University, York, UK.

Contributors: Laura Manea contributed to the acquisition, analysis and interpretation of the data and drafted the manuscript. Simon Gilbody and Dean McMillan contributed to the concept and design of the study and the interpretation of data and revised the manuscript for important intellectual content. All of the authors approved the final version of the manuscript submitted for publication.

Funding: No funding was received for this study.