Identifying depression with the PHQ-2: A diagnostic meta-analysis

(Short title: Identifying depression with the PHQ-2)

Laura Manea MMedSci MRCPsych[1] , Simon Gilbody DPhil FRCPsych FRSA[1], Catherine Hewitt PhD[2], Alice North[2], Faye Plummer MSc[2], Rachel Richardson MA MBA[2], Brett D. Thombs PhD[3], Bethany Williams[2], Dean McMillan PhD[1*]

[*]Corresponding author

[1]Hull York Medical School and Department of Health Sciences, University of York

[2]Department of Health Sciences, University of York

Correspondence details:

Dean McMillan PhD, DClinPsy, Hull York Medical School and Department of Health Sciences, ARRC Building, University of York, YO10 5DD

Email: dean.mcmillan@york.ac.uk

Abstract: 335 words

Article text: 4,326

Key words: major depression, screening, diagnostic accuracy, PHQ-2, ultra-brief screening instruments, diagnostic meta-analysis.

Pages: 60

Abstract

Background

There is interest in the use of very brief instruments to identify depression because of the advantages they offer in busy clinical settings. The PHQ-2, consisting of two questions relating to core symptoms of depression (low mood and loss of interest or pleasure), is one such instrument.

Method

A systematic review was conducted to identify studies that had assessed the diagnostic performance of the PHQ-2 to detect major depression. Embase, MEDLINE, PsychINFO and grey literature databases were searched. Reference lists of included studies and previous relevant reviews were also examined. Studies were included that used the standard scoring system of the PHQ-2, assessed its performance against a gold-standard diagnostic interview and reported data on its performance at the recommended (≥3) or an alternative cut-off point (≥2). After assessing heterogeneity, where appropriate, data from studies were combined using bivariate diagnostic meta-analysis to derive sensitivity, specificity, likelihood ratios and diagnostic odds ratios.

Results

21 studies met inclusion criteria totalling N=11,175 people out of which 1,529 had major depressive disorder according to a gold standard. 19 of the 21 included studies reported data for a cut-off point of ≥3. Pooled sensitivity was 0.76 (95% CI = 0.68 – 0.82), pooled specificity was 0.87 (95% CI = 0.82 – 0.90). However there was substantial heterogeneity at this cut-off (I^2 = 81.8%). 17 studies reported data on the performance of the measure at cut-off point ≥2. Heterogeneity was I^2 = 43.2% pooled sensitivity at this cut-off point was 0.91 (95% CI = 0.85–0.94), and pooled specificity was 0.70 (95% CI = 0.64–0.76).

Conclusion

The generally lower sensitivity of the PHQ-2 at cut-off ≥3 than the original validation study (0.83) suggests that ≥2 may be preferable if clinicians want to ensure that few cases of depression are missed. However, in situations in which the prevalence of depression is low, this may result in an unacceptably high false-positive rate because of the associated modest specificity. These results, however, need to be interpreted with caution given the possibility of selectively reported cut-offs.

Introduction

Depression is common and disabling, but its management is suboptimal in primary and secondary care [1]. Screening has been proposed as a solution to improving depression care, but the value of routine screening and case finding procedures to detect depression has not been proven [1,2]. Some national guidelines recommend it in primary care [3], whereas others do not.[4,5]

Recently there has been an increased interest in the potential of using very brief instruments to identify patients with major depression, because of the advantages they may offer in busy clinical settings in which time is limited [6]. One such very brief screening measure for depression is the two-item Patient Health Questionnaire (PHQ-2) [7], an abbreviated version of the widely used PHQ-9 [8]. It is comprised of the first two questions of the PHQ-9, which reflect the core symptoms of depression (low mood, loss of interest/pleasure). The original validation study of the PHQ-2 provided preliminary evidence that it may be an effective screen for depression [7]. In that study, a cut-off point of ≥3 (out of a possible score of 6) had a sensitivity of 0.83 and a specificity of 0.90 to identify major depression in a sample of 580 primary and secondary care patients, although this included only 41 patients with major depression, a small number for estimating diagnostic accuracy. This contrasts favourably with sensitivity of 0.88 and specificity of 0.88 in the nine-item PHQ-9 among the same patients [8].

A previous systematic review of the diagnostic properties of the PHQ-2 identified only a small number of studies (N = 3) that had examined the diagnostic performance of the PHQ-2 [9]. The review concluded that no recommendations could be made about the PHQ-2 without further validation studies across a range of clinical settings and populations. The authors of the review, however, did suggest that preliminary evidence suggested that the PHQ-2 could be a brief, yet accurate tool. Since that initial review the PHQ-2 has been much more widely evaluated in primary studies, but there is not an updated systematic review. The current systematic review aims to evaluate the current evidence base for the PHQ-2 to identify patients with major depression.

Methods

Literature search

We searched Embase, MEDLINE, PsycINFO and grey literature databases (OIASTER, OpenGrey, ZETOC) from inception to August 2014. The search terms used for Embase, Medline and PsycINFO are given in appendix 1. The terms were adapted as necessary for the grey databases. In addition, we examined the reference lists of all included studies and previous relevant reviews, including reviews of the PHQ-9 [9-12] and a review of ultra-brief screening instruments for depression [6].

Study selection

A pre-piloted coding manual outlining a priori inclusion-exclusion criteria along with operational definitions of each was developed. *Population:* Any population or setting was included. *Instrument:* We included studies that used the PHQ-2 scored in the standard way (each item scored 0-3 and summed to give a total score between 0 and 6). Studies that used atypical methods of scoring the PHQ-2 (e.g., scored as positive if either item was scored as two or above) were excluded. *Comparison (reference standard):* The accuracy of the PHQ-2 had to be assessed against a recognised gold-standard instrument for the diagnosis of either Diagnostic and Statistical Manual (DSM) or International Classification of Disease (ICD) criteria for major depression. Studies that used other reference standards, such as unaided clinician diagnosis or scores above a cut-off point on another self-report instrument, were excluded. Studies were also excluded if the target diagnosis was not major depression (e.g., any depressive disorder). *Outcome:* Studies had to report sufficient information to calculate a 2*2 contingency table for the cut-off point ≥3 recommended by the original validation study or the lower, alternative cut-off recommended by some studies (≥2). *Study design:* Any design. *Additional criterion:* Studies were excluded if the sample overlapped with that used in another included study. Citations with overlapping samples were examined to establish whether they contained information relevant to the research question that was not contained in the included report. We included in the review the study that had the larger sample or, if the samples were the same size, the study that provided all the details required for tis review. No restrictions were made in terms of publication status, publication year or language.

All identified citations were first assessed on the basis of title and abstract. At this stage, the inclusion-exclusion criteria were interpreted liberally; if there was doubt about whether a citation met the criteria it was included. Full paper copies of those that passed this first sift were obtained and examined in detail against the inclusion-exclusion criteria. Studies that met this second sift were included in the systematic review. Where necessary authors were contacted to provide further clarification or to obtain additional information.

Data extraction

We extracted the following data to a pre-piloted, standardised form: sample characteristics (country, setting, age, gender), sample size and percentage with major depression according to the gold standard, information on the PHQ-2 (method of administration, cut-offs reported, language), and details of the reference standard. In addition, we calculated cell Ns of the 2*2 tables at cut-offs ≥2 and ≥3. Again, where necessary authors were contacted to provide clarification.

Quality assessment

Quality assessment was conducted at the study level and used criteria based on the QUADAS-2 (the revised tool for the Quality Assessment of Diagnostic Accuracy Studies) [13]. 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 The QUADAS-2 guidelines require that it is adapted for each specific

review; this can involve adding or omitting questions and providing clarification about how specific questions are to be rated. We retained all of the risk of bias signaling questions and applicability questions, for which we developed specific guidance on coding in the form of a brief field guide. For the signaling question 'Is the reference standard likely to correctly classify the target condition?' we operationalised this as whether the researchers who conducted the gold standard interview had received appropriate training. For the signaling question 'Was there an appropriate interval between the index test and reference standard?' we defined an appropriate interval as less than two weeks in keeping with how this item has been applied in previous diagnostic test accuracy studies of depression [14].

We added four additional questions that were applied to studies using translated versions of the PHQ-2 and reference test. For translations of the PHQ-2, we asked whether appropriate translation methods were used and whether psychometric properties of the translated version were reported. The same two questions (appropriate translation, psychometric properties) were also applied to any translated version of the reference test.

Data analysis and synthesis

Sensitivity, specificity, positive and negative likelihood ratios and diagnostic odds ratios along with their associated 95% confidence intervals were calculated for cut-off points ≥2 and ≥3. Heterogeneity was assessed using I^2 for the diagnostic odds ratio, an estimate of the proportion of

study variability that is due to between-study variability rather than sampling error. We considered values of ≥50% to indicate substantial heterogeneity [15]. Where heterogeneity was not substantial we used bivariate diagnostic meta-analyses to generate pooled estimates of sensitivity and specificity. Summary Receiver Operating Characteristics (sROC) were calculated to produce 95% confidence interval ellipses within ROC space.

Where substantial heterogeneity was identified, we conducted pre-planned subgroup analyses based on clinical setting. We further explored possible reasons for heterogeneity by conducting pre-planned meta-regressions of key descriptive variables and the quality assessment criteria [15].

We attempted to limit publication bias by searching a range of grey literature databases. The potential for selective outcome reporting bias related to the reporting of results for some but not other cut-off points is explored in the discussion section.

Bayesian nomograms were generated to examine the performance of the PHQ-2 at different prevalence estimates.

Results

The initial search identified 1054 unique citations (2882 citations before de-duplication). 59 of these citations met initial inclusion criteria and were selected for further screening of the full article. 21 of the 59 met final stage inclusion criteria[7,16-34].

The remaining 38 were excluded for the following reasons: screening instrument was not the PHQ-2 (N = 9), PHQ-2 was scored in a non-standard way (N = 7), reference standard was not a recognised gold-standard instrument (N = 7), reference standard diagnosis was not solely major depression (N = 3), study reported insufficient information to calculate a 2*2 table for at least one of the cut-off points (N = 2), and overlap in samples with included studies (N = 7). Two additional citations were excluded because we were unable to obtain further information from the authors to establish whether they met inclusion criteria. Finally, one study was excluded, as all included patients were known to have depression and would, thus, not be screened in practice. The selection of studies is summarised in the PRISMA flowchart [35] in figure 1 and further details about the reasons for exclusion are given in appendix 2.

FIGURE 1

ABOUT HERE

Overview of included studies

Table 1 summarises the characteristics of the included studies. Three studies used general primary care samples [7,16,36], with a further one focused on older adults in primary care [37]. One study focused on patients with epilepsy, but recruited these from primary care [38]. A further three studies used a combination of a primary care setting and another setting, such as outpatient clinics [24,39,40]. Eight studies recruited from hospital- or out-patient-based medical specialties [19,20,22,23,29,32,41,42]. Of the remainder, one recruited from a community-drug treatment service [21], one from a community-based aging service [28], one from a research institute focusing on adolescents [27] and two from community settings (students) [31,33].

--

TABLE 1

ABOUT HERE

All of the studies apart from two [27,31] had working age or older adult samples. In the majority of studies, there were markedly more females than males or the samples were entirely female. The proportion of the sample that met reference standard criteria for major depression ranged from 2% [19] to 61.2% [21]. Some of the studies had a high prevalence of depression because the study design over-sampled people with positive PHQ-2 scores for administration of the reference standard. [27,32,38].

Six studies stated that a self-report version of the PHQ-2 was used [7,21,22,24,31,36,37]. In one study it was administered over the telephone [27] and in four studies it was administered face to face [17,20,23,33]; the remaining studies did not clearly state the method of administration. Translated versions of the PHQ-2 were used in ten studies [17,19-21,23,24,31,33,36,40], including Brazilian, Chinese, Dutch, Japanase and German versions.

Quality assessment

Table 2 summarises the results of the quality assessment using QUADAS-2. The studies varied in quality. Only two of the studies were judged to be at a low risk of bias across all of the domains [16,36]. One of these studies [36], however, was the only one not to meet all of the applicability criteria. The reference standard in Zuithoff et al. [36] assessed major depression over a one-year time-frame, so, unlike the PHQ-2, is not assessing current depression. This may have lowered the observed accuracy of the PHQ-2 in that study. A number of studies had high

prevalence rates of depression because the studies use a design in which participants who are at an increased risk of depression (e.g. those scoring above a threshold on the PHQ-2) were more likely to be given the reference standard [27,32,38].

TABLE 2

ABOUT HERE

Narrative overview of diagnostic performance

Table 3 summarises the test accuracy characteristics of the PHQ-2 at the standard cut-off point of ≥3; table 4 gives the same data for the alternative cut-off point of ≥2.

TABLES 3 & 4

ABOUT HERE

Nineteen studies reported the performance of the PHQ-2 at cut-off point ≥3. At this cut-off, sensitivity ranged from 0.39 [30] to 1 [19] and specificity from 0.59 [29] to 1 [38]. Five studies, one of which was the original validation study, were conducted in primary care. Of these, one study focused solely on people with epilepsy [38] so was not considered a general primary care sample.

Seventeen studies reported details of the performance of the PHQ-2 at cut-off point ≥2 (see table 4). The distinction between the performance of the PHQ-2 in the original validation study and the other studies was less marked than at cut-off point ≥3, though for those studies in which a diagnostic odds ratio could be calculated, the value was higher in the original validation studies than the subsequent studies.

Diagnostic meta-analyses

An initial diagnostic meta-analysis was run including all 19 studies reporting the performance of the PHQ-2 at cut-off point ≥3. Pooled sensitivity was 0.76 (95% CI 0.68 – 0.82), pooled specificity 0.87 (95% CI 0.82 – 0.90), pooled positive likelihood ratio 6.02 (95% CI 4.44 – 8.18), pooled negative likelihood ratio 0.27 (95% CI 0.20 – 0.36) and pooled diagnostic odds ratio 22.20 (95% CI 14.00 -35.19).

One of the possible reasons for heterogeneity is the various clinical settings in which the PHQ-2 has been validated. On a priori grounds we conducted subgroup analyses to examine the diagnostic performance of the PHQ-2 in similar clinical settings. As described above, of the five

primary care studies one focused solely on people with epilepsy so could not be considered a general primary care sample and was excluded [38]. A diagnostic meta-analysis was conducted for the remaining four primary care studies [7,16,36,37]; however, heterogeneity remained substantial (I^2 = 67.7%). Pooled sensitivity was 0.64 (95% CI = 0.46 – 0.78) and pooled specificity was 0.91 (95% CI = 0.89 – 0.93). Six studies that reported cut-off point 3 were conducted in secondary care. [19,23,29,32,41,42]. Pooled sensitivity was 0.74 (95% CI = 0.57 – 0.86) and pooled specificity was 0.85 (95% CI = 0.74 – 0.91). Heterogeneity was high for this group as well (I^2 = 73.3%).We did not identify a sufficient number of studies (minimum of four studies for a diagnostic meta-analysis to be carried out in STATA) using a comparable clinical setting to conduct further subgroup analyses for other settings.

We conducted a meta-regression to further explore other possible sources of heterogeneity. Descriptive variables (setting, age, proportion female, language) were examined as predictors as were the individual quality criteria. P values were calculated using STATA metareg hand written command. None was significant at $p < 0.05$.

As previously mentioned, in one study (Zuithoff et al. [36]) the reference standard assessed major depression over a one-year time-frame. Excluding this study from the meta-analyses did not significantly alter the pooled results.

An initial diagnostic meta-analysis was run for the 17 studies reporting the performance of the PHQ-2 at cut-off point ≥2. Pooled sensitivity was 0.91 (95% CI = 0.85 – 0.94) and pooled specificity was 0.70 (95% CI = 0.64 – 0.76) (see figure 2 for sROC). Heterogeneity was moderate (I^2 = 43.5%). When the analysis was rerun for the four primary care studies [7,16,36,37], this gave a pooled sensitivity of 0.84 (95% CI = 0.80 – 0.88) and pooled specificity of 0.76 (95% CI = 0.74 – 0.79) (see figure 3 for sROC). Heterogeneity was still moderate (I^2 = 42.3%). Five studies that reported cut-off point of 2 were conducted in secondary care settings [19,20,22,23,41]. Pooled sensitivity was 0.84 (95% CI = 0.68 – 0.92) and pooled specificity was 0.76 (95% CI = 0.65 – 0.85).

Descriptive variables (setting, age, proportion female, language) and the individual quality criteria were not identified as sources of heterogeneity in meta-regression analyses for the studies that reported cut-off point 2 ($p > 0.05$).

FIGURES 2 & 3

ABOUT HERE

Figure 4 uses the pooled sensitivity and specificity at cut-off ≥2 to estimate the performance of the PHQ-2 at this cut-off point as prevalence varies. The diagonal line in blue represents the prevalence of depression. The probability that a person is depressed according to the gold standard given a positive score is represented by the red line; the probability that a person is depressed given a negative score is represented by the green line.

FIGURES 4 & 5

ABOUT HERE

Discussion

The original validation study of the PHQ-2 recommended a cut-off point of ≥3 on the basis of a sensitivity of 0.83 and specificity of 0.90 [7]. This systematic review suggests that the accuracy of the PHQ-2 in identifying major depression is lower than that reported in the original study at this cut-off point. In general, sensitivity was lower than that reported in the original validation study [7]. This, however, was not necessarily

linked to the other studies reporting higher specificity, as may be expected given that sensitivity and specificity are inversely related. As a result, for those studies for which a diagnostic odds ratio could be calculated, with the exception of two studies [23,39], all had a lower diagnostic odds ratio than the figure of 43.6 (95% CI = 18.8 – 101) calculated for Kroenke et al. (2003) [7]. There was substantial heterogeneity at ≥3, which makes difficult the interpretation of pooled sensitivity and specificity. For the primary care studies, the sensitivity was substantially lower than Kroenke et al. (2003) [7] (0.64 compared to 0.83 in the original validation study) and this was paired with broadly comparable levels of specificity. (0.91 compared to 0.90).

Lowering the cut-off point will increase sensitivity. Pooled sensitivity at the cut-off point of ≥2 was 0.91 (95% CI = 0.85 – 0.94), which is higher than the sensitivity reported in the original validation study at cut-off point ≥3. This, however, would come at the cost of lowered specificity given its inverse relationship with sensitivity. At a cut-off point of ≥2 pooled specificity was 0.70 (95% CI = 0.64 – 0.76). The pooled values for the primary care samples were broadly comparable (pooled sensitivity = 0.84, 95% CI = 0.80 – 0.88; pooled specificity = 0.76, 95% CI = 0.74 – 0.79).

While the lowering of the cut-off point may limit the number of people that would be missed by the screen, it is unclear whether the level of false positives generated by this strategy would be acceptable to clinicians. The extent to which this would be a problem depends on the prevalence of depression in which the screen is being used and the cost and availability of strategies to further assess those who score positively on the initial screen.

As prevalence falls, the proportion of people who score positively but who are not depressed will increase. Prevalence estimates from the studies reported here vary substantially, though for some of the higher estimates this is likely to be related to sampling strategies that over-selected people who were likely to be depressed [27,32,38]. Some idea of the value of using a cut-off point of ≥2 can be gained by using the pooled sensitivity and specificity values to estimate the proportion of people scoring ≥2 who were in fact depressed according to the reference standard at different prevalence estimates (see figure 4). For illustrative purposes, prevalence values of 5, 15 and 25% are discussed. On the basis of the pooled sensitivity and specificity values, at a 5% prevalence of depression approximately 14% of people who scored at ≥2 would be depressed according to the gold standard; at 15% prevalence the value becomes approximately 37% and at 25% prevalence the value would be 51%. The pooled sensitivity and specificity of the primary care studies at this cut-off point gives similar results (5% prevalence: 16%; 15% prevalence: 38%; 25% prevalence 54%) (see figure 5). This analysis assumes that no patients are being treated for depression, which is perhaps an unrealistic

assumption. About half of patients are recognized without screening and in primary care and a large number are already treated. However the studies do not present sufficiently detailed data to re-run the analyses for people not known to be depressed.[43]

At the lower estimates of prevalence, this cut-off point may generate too high a proportion of people scoring positively who are not depressed to make it a useful clinical tool. This suggests that it may be of limited use as a case-finding instrument, in which all people presenting to a service, such as a general practitioner surgery, are opportunistically screened, because in such a context the prevalence is likely to be low. As the prevalence increases, however, it may become useful. This suggests that the PHQ-2 at a cut-off point of ≥2 may be of use in screening situations in which a group known to be at high risk of depression is targeted for screening, because of the increased prevalence of depression. There are, however, a number of caveats to this conclusion. First, the studies reviewed here typically used it in a general screening context; evaluation in selective contexts would be needed to confirm it performance in these situations. Secondly, as already mentioned, the studies reviewed do not distinguish between those people who are already known to services to be depressed and those who are depressed but not known. The aim of selective screening would be to identify cases that are not already known to clinical services. The prevalence of previously unknown depression will be lower than the overall depression prevalence, which may again limit the value of any identification tool. It is also unclear how the different context of identifying only previously unidentified depression would affect the diagnostic characteristics of the measure. Thirdly, the value of a screening tool cannot be assessed solely on the basis of its sensitivity and specificity, but can only be assessed as part of a wider

evaluation that examines the effectiveness and cost-effectiveness of not only screening, but the consequences of screening in terms of treatment and the outcome of that treatment. [5]

While this cut-off point may have some limitations in identifying people likely to have depression when there is a low prevalence of depression, given the high false positive rate, the negative likelihood ratios for this cut-off point suggest that those people who are predicted to be not depressed according to this cut-of point are unlikely to be depressed, particularly when the prevalence of depression is low. The PHQ-2 at ≥2, therefore, may have value in ruling out depression. Figure 4 illustrates this for the pooled sensitivity and specificity. If the pooled sensitivity and specificity values are used, at 5% prevalence approximately 99% of people scoring below the cut-off would not be depressed; at 15% the figure is 97% and at 25% the figure is 94%. The corresponding figures based on the primary care pooled estimates of sensitivity and specificity are 99% (5% prevalence), 96% (15% prevalence) and 93% (25% prevalence) (see figure 5).

It is important to note that the results of this meta-analysis do not apply to the Whooley questions (also known as the 'yes/no' PHQ-2). The Whooley questions are often confused with, and referred to as, the PHQ-2. However, the relatively poor sensitivity and specificity reported for the PHQ-2 in this study does not apply to the Whooley questions. A recent diagnostic meta-analysis of the Whooley questions has shown that the Whooley qestions appear to be more sensitive and efficient for screening purposes.

Limitations

Although we sought to review grey literature databases, we cannot rule out the possibility of publication bias. Study selection and data extraction were performed by one author, which may have also introduced bias.

Three studies [27,32,38] used a design in which participants who were more likely to be depressed were also more likely to be given the reference standard, which may have introduced a partial verification bias. The QUADAS-II assessment identified variability in study quality, with only a small number of studies rated as at low risk of bias across all domains. Variations in study quality, however, did not appear to be related to outcome according to the meta-regression for cut-off point ≥3.

There was some lack of detail in the reporting of studies, which made it difficult to assess some of the QUADAS-2 criteria. This was particularly the case for the reporting of whether the reference standard was conduced blind to the PHQ-2. Future studies should make clear statements about the blinding of the reference standard and more generally ensure that the method is reported in sufficient detail to assess the standard QUADAS-2 criteria.

Some studies may have selectively reported cut-off points - the studies that reported the two cut-off points (2 and 3) varied. It is possible that there is a relationship between the observed performance of the PHQ-2 at a particular cut-off point and the likelihood that it is reported for a particular study. Future studies should report the performance of the PHQ-2 at all available cut-off points to protect against the possibility of selective outcome reporting. Some studies reported details of sensitivity and specificity but were excluded because we were unable to identify the additional information required to calculate the 2*2 tables that permit the calculation of the full range of accuracy statistics. Future studies should also report sufficient information to ensure that a 2*2 table can be reconstructed form the information reported. As described above, the role of screening is to identify previously unknown cases, yet typically the studies identified in this review do not differentiate between previously known and previously unknown cases. It is not clear what impact restricting the analysis to previously unknown cases would have on sensitivity and specificity, but such an approach would necessarily reduce the prevalence of depression, which may affect whether the instrument is likely to be useful in a particular clinical context. Future validation studies should seek to report the diagnostic performance of the PHQ-2 in identifying previously unknown cases.

The pooled estimates should be interpreted with caution given the high level of heterogeneity. Although I^2 may exaggerate heterogeneity in DTA studies, there is no clear guidance available on the best way to manage this.

Another interesting finding of this review is the relatively small number of validation studies of the PHQ-2 compared to the number of validation studies of the PHQ-9, which incorporates the PHQ-2. A recent meta-analysis of the PHQ-9 has identified 36 validation studies and most of these do not specifically report the psychometric properties of the PHQ-2.

Conclusion

In screening situations, reasonably high sensitivity is often required to ensure that the screening process misses few people with the diagnosis. The original validation study of Kroenke et al. (2003) [7] reported sensitivity of 0.83 at a cut-off point of ≥3, but a number of subsequent studies have tended to report somewhat lower sensitivity at this cut-off point. If sensitivity comparable to that reported in the original validation study is required in a screening situation, then the lower cut-off point may be needed to ensure sufficiently high sensitivity. However, the associated specificity value at this cut-off point is modest, which may limit the usefulness of the PHQ-2 at this cut-off point to identify people likely to be depressed when the prevalence of depression is low.

Acknowledgements

We would like to thank the authors of both the included and excluded studies for their help in answering our questions about their studies.

Figure legend

Figure 1: PRISMA flow chart PHQ-2

Figure 2: sROC for cut-off point ≥3 (19 studies)

Figure 3: sROC for cut-off point ≥2 (17 studies)

Figure 4: Performance of PHQ-2 at ≥2 using pooled sensitivity and specificity at different prevalence estimates

Figure 5: Performance of PHQ-2 at ≥2 using pooled sensitivity and specificity at different prevalence estimates in primary care studies

References

1. Gilbody S, Sheldon T, House A (2008) Screening and case-finding instruments for depression: A meta-analysis. Canadian Medical Association Journal 178: 997-1003.

2. Thombs BD, Coyne JC, Cuijpers P, de Jonge P, Gilbody S, et al. (2012) Rethinking recommendations for screening for depression in primary care. Canadian Medical Association Journal 184: 413-418.

3. U.S. Preventive Services Task Force (2009) Screening for depression in adults: US Preventive Services Task Force recommendation statement. Annals of Internal Medicine 151: 784-792.

4. Joffres M, Jaramillo A, Dickinson J, Lewin G, Pottie K, et al. (2013) Recommendations on screening for depression in adults. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 185: 775-782.

5. Allaby M (2010) Screening for depression: A report for the UK National Screening Committee. UK National Screening Committee.

6. Mitchell AJ, Coyne JCJ (2007) Do ultra-short screening instruments accurately detect depression in primary care? A pooled analysis and meta-analysis of 22 studies. British Journal of General Practice 57: 144-151.

7. Kroenke K, Spitzer RL, Williams JB (2003) The Patient Health Questionnaire-2: validity of a two-item depression screener. Medical Care 41: 1284-1292.

8. Kroenke K, Spitzer RL, Williams JBW (2001) The PHQ-9: validity of a brief depression severity measure. . Journal of General Internal Medicine 16: 606-613.

9. Gilbody S, Richards D, Brealey S, Hewitt C (2007) Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. Journal of General Internal Medicine 22: 1596-1602.

10. Wittkampf KA, Naeije L, Schene AH, Husyer J, van Weert HC (2007) Diagnostic accuracy of the mood module of the Patient Health Questionnaire: A systematic review. General Hospital Psychiatry 29: 388-395.

11. Kroenke K, Spitzer RL, Williams JBW, Lowe B (2010) The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: A systematic review. General Hospital Psychiatry 32: 345-359.

12. Manea L, Gilbody S, McMillan D (2012) Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. Canadian Medical Association Journal 184: E191-E196.

13. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, et al. (2011) QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Annals of Internal Medicine 155: 529-536.

14. Mann R, Hewitt CE, Gilbody SM (2009) Assessing the quality of diagnostic studies using psychometric instruments: Applying QUADAS. Social Psychiatry and Psychiatric Epidemiology 44: 300-307.

15. Centre for Reviews and Dissemination (2009) Systematic Reviews: CRD's guidance for undertaking reviews in health care. York: University of York.

16. Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, et al. (2010) Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. Annals of family medicine 8: 348-353.

17. Chagas MH, Crippa JA, Loureiro SR, Hallak JE, Meneses-Gaya C, et al. (2011) Validity of the PHQ-2 for the screening of major depression in Parkinson's disease: two questions and one important answer. Aging & mental health 15: 838-843.

18. de Lima Osorio F, Mendes AV, Alexandre Crippa J, Loureiro SR (2009) 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. [References]. Perspectives in Psychiatric Care 45: 216-227.

19. Osorio F, Carvalho A, Fracalossi T, Crippa J, Loureiro E (2012) Are two items sufficient to screen for depression within the hospital context. International Journal of Psychiatry in Medicine 44: 141-148.

20. de Man-van Ginkel JM, Gooskens F, Schepers VP, Schuurmans MJ, Lindeman E, et al. (2012) Screening for poststroke depression using the patient health questionnaire. Nursing Research 61: 333-341.

21. Delgadillo J, Payne S, Gilbody S, Godfrey C, Gore S, et al. (2011) How reliable is depression screening in alcohol and drug users? A validation of brief and ultra-brief questionnaires. Journal of Affective Disorders 134: 266-271.

22. Fiest KM, Patten SB, Wiebe S, Bulloch AGM, Maxwell CJ, et al. (2014) Validating screening tools for depression in epilepsy. Epilepsia 55: 1642-1650.

23. Inagaki M, Ohtsuki T, Yonemoto N, Kawashima Y, Saitoh A, et al. (2013) Validity of the Patient Health Questionnaire (PHQ)-9 and PHQ-2 in general internal medicine primary care at a Japanese rural hospital: A cross-sectional study. General Hospital Psychiatry 35: 592-597.

24. Lowe B, Kroenke K, Grafe K (2005) Detecting and monitoring depression with a two-item questionnaire (PHQ-2). Journal of Psychosomatic Research 58: 163-171.

25. Margrove K, Mensah S, Thapar A, Kerr M (2011) Depression screening for patients with epilepsy in a primary care setting using the Patient Health Questionnaire-2 and the Neurological Disorders Depression Inventory for Epilepsy. Epilepsy and Behavior 21: 387-390.

26. Phelan E, Williams B, Meeker K, Bonn K, Frederick J, et al. (2010) A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. BMC family practice 11.

27. Richardson LP, McCauley E, Grossman DC, McCarty CA, Richards J, et al. (2010) Evaluation of the patient health questionnaire-9 item for detecting major depression among adolescents. Pediatrics 126: 1117-1123.

28. Richardson TM, He H, Podgorski C, Tu X, Conwell Y (2010) Screening depression aging services clients. American Journal of Geriatric Psychiatry 18: 1116-1123.

29. Smith MV, Gotman N, Lin H, Yonkers KA (2010) Do the PHQ-8 and the PHQ-2 accurately screen for depressive disorders in a sample of pregnant women? General Hospital Psychiatry 32: 544-548.

30. Thombs BD, Ziegelstein RC, Whooley MA (2008) Optimizing detection of major depression among patients with coronary artery disease using the patient health questionnaire: Data from the heart and soul study. Journal of General Internal Medicine 23: 2014-2017.

31. Tsai FJ, Huang YH, Liu HC, Huang KY, Liu SI (2014) Patient health questionnaire for school-based depression screening among Chinese adolescents. Pediatrics 133.

32. Williams LS, Brizendine EJ, Plue L, Bakas T, Tu W, et al. (2005) Performance of the PHQ-9 as a screening tool for depression after stroke. Stroke 36: 635-638.

33. Zhang Y, Ting R, Lam M, Lam J, Nan H, et al. (2013) Measuring depressive symptoms using the Patient Health Questionnaire-9 in Hong Kong Chinese subjects with type 2 diabetes. Journal of Affective Disorders 151: 660-666.

34. Zuithoff NP, Vergouwe Y, King M, Nazareth I, van Wezep MJ, et al. (2010) The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a crosssectional study. BMC family practice 11.

35. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Journal of Clinical Epidemiology 62: 1006-1012.

36. Zuithoff NP, Vergouwe Y, King M, Nazareth I, van Wezep MJ, et al. (2010) The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a crosssectional study. BMC family practice 11: 1-7.

37. Phelan E, Williams B, Meeker K, Bonn K, Frederick J, et al. (2010) A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. BMC family practice 11: 1-9.

38. Margrove K, Mensah S, Thapar A, Kerr M (2011) Depression screening for patients with epilepsy in a primary care setting using the Patient Health Questionnaire-2 and the Neurological Disorders Depression Inventory for Epilepsy. Epilepsy & Behavior 21: 387-390.

39. De Lima Osorio F, Vilela Mendes A, Crippa JA, Loureiro SR (2009) 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. Perspectives in Psychiatric Care 45: 216-227.

40. Liu SI, Yeh ZT, Huang HC, Sun FJ, Tjung JJ, et al. (2011) Validation of Patient Health Questionnaire for depression screening among primary care patients in Taiwan. Comprehensive Psychiatry 52: 96-101.

41. Chagas MHN, Crippa JAS, Loureiro SR, Hallak JEC, de Meneses-Gaya C, et al. (2011) Validity of the PHQ-2 for the screening of major depression in Parkinson's disease: Two questions and one important answer. [References]. Aging & mental health 15: 838-843.

42. Thombs BD, Ziegelstein RC, Whooley MA (2008) Optimizing detection of major depression among patients with coronary artery disease using the Patient Health Questionnaire: Data from the Heart and Soul Study. [References]. Journal of General Internal Medicine 23: 2014-2017.

43. Thombs BD, Arthurs E, El-Baalbaki G, Meijer A, Ziegelstein RC, et al. (2011) Risk of bias from inclusion of patients who already have diagnosis of or are undergoing treatment for depression in diagnostic accuracy studies of screening tools for depression: systematic review. British Medical Journal 343.

Table 1: Descriptive characteristics of the included studies

Study	Sample characteristics (Country, setting, age, sex)	Sample size and % depressed	PHQ-2 characteristics	Diagnostic standard
Arroll et al. (2010)	Country: New Zealand Setting: Primary care Age (yrs): Av. = 49 (range = 17-99) Female: 61%	N = 2642 Depressed: 6.2%	Administration: Not stated Language: English	DSM-IV CIDI
Chagas et al. (2011)	Country: Brazil Setting: Movement disorders outpatient clinic Age (yrs): M = 71.09 (sd = 12.62) Female: 53%	N = 110 Depressed: 25.5%	Administration: Neurologist administered Language: Brazilian	DSM-IV SCID
De Lima Osorio et al. (2009)	Country: Brazil Setting: Gynaecology and General Practice Age (yrs): 48% < 30 Female: 100%	N = 177 Depressed: 34%	Administration: Not stated Language: Brazilian Portuguese	DSM-IV SCID
De Lima Osorio et al. (2012)	Country: Brazil Setting: General hospital Age (yrs): M = 49 (SD =12.4)	N = 100 Depressed: 2%	Administration: Not stated Language: Brazilian Portuguese	DSM-IV SCID

	Female: 39%			
De Man-van Ginkel et al. (2012)	Country: Netherlands Setting: Stroke patients Age (yrs): M = not specified Female: % not specified	N = 164 Depressed: 12.2%	Administration: Face to face Language: Unclear (?Dutch and English)	CIDI
Delgadillo et al. (2011)	Country: UK Setting: Community drug treatment service Age (yrs): M = 35 (range: 23-54) Female: 23%	N = 103 Depressed: 61.2%	Administration: Self-report (assistance if required) Language: English	ICD-10 CIS-R
Fiest et al. (2014)	Country: Canada Setting: Secondary care (epilepsy clinic) Age (yrs): M = 40.3 (range: 18.2 – 78.1) Female: 51.4 %	N= 185 Depressed: 14.6%	Administration: Self-report Language: English	DSM IV/V SCID
Inagaki et al. (2013)	Country: Japan Setting: Secondary care (general medical clinic) Age (yrs): M = 73.5 (SD 12.3)	N= 104 Depressed: 7.4%	Administration: Face to face Language: Japanese	MINI

	Female: 59.3 %			
Kroenke et al. (2003)	Country: US Setting: Primary care Age (yrs): Primary: M = 46 Female: Primary = 66%	N = 580 Depressed: 7.1%	Administration: Self-report Language: English	DSM-III-R PRIME-MD
Liu et al. (2011)	Country: Taiwan Setting: Community-based primary care and hospital-based family physician clinics Age (yrs): Not reported Female: % not reported	N = 1532 Depressed: 3.3%	Administration: Not stated Language: Chinese	DSM-IV SCAN
Lowe et al. (2005)	Country: Germany Setting: Outpatient clinics and family practices Age (yrs): M = 42.0 (sd = 13.8) Female: 67.5%	N = 520 Depressed: 13.7%	Administration: Self-report Language: German	DSM–IV SCID
Margrove et al. (2011)	Country: UK Setting: Diagnosis of epilepsy in primary care Age (yrs): M = 49 (sd = 16)	N = 52 Depressed: 48.1%	Administration: Self-report Language: English	DSM-IV SCID

	Female: 49.8%			
Phelan et al. (2010)	Country: US Setting: Older adults in primary care clinics Age (yrs): M = 78 (sd = 7) Female: 62%	N = 69 Depressed: 12%	Administration: Self-report (assistance if required) Language: English	DSM-IV SCID
Richardson et al. (2010)	Country: US Setting: Group Health Research Institute Age (yrs): M = 15.3 (sd = 1.1) Female: 60%	N = 444 Depressed: 54.5%	Administration: Telephone administered Language: English	DSM-IV DISC
Richardson et al. (2010)	Country: US Setting: Community-based aging services agency Age (yrs): M = 76.5 (sd = 9.2) Female: 68.5%	N = 378 Depressed: 26.7%	Administration: Unclear Cut-offs: ≥1 to 6 Language: English	DSM-IV SCID
Smith et al. (2010)	Country: US Setting: Obstetrical settings	N = 213 Depressed: 6.1%	Administration: Not stated Language: English	DSM-IV CIDI

	Age (yrs): Depressed: 29.31 (sd = 5.98) Non depressed: 28.87 (sd = 6.72) Female: 100%			
Thombs et al. (2008)	Country: US Setting: Outpatients with coronary heart disease Age (yrs): M = 67 (sd = 11) Female: 18%	N = 1024 Depressed: 22%	Administration: Not stated Language: English	DSM C-DIS
Tsai et al. (2014)	Country: Taiwan Setting: Community (high-schools) Age (yrs): M = 16.9 (sd = 0.6) Female: 59.6%	N= 165 Depressed 10%	Administration: Self-report Language: Chinese	DSM K-SADS-E
Williams et al. (2005)	Country: US Setting: Inpatient stroke Age (yrs): 42% < 60 Female: 51%	N = 316 Depressed: 34%	Administration: Not stated Language: English	DSM-IV SCID
Zhang et al. (2013)	Country: China	N = 959 Depressed: 8.8%	Administration: Face to face Language: Chinese	DSM-IV SCID

	Setting: Community (university students) Age (yrs): M = 21.45 (sd = 1.04) Female: 54.3%			
Zuithoff et al. (2010)	Country: Netherlands Setting: Primary care Age (yrs): M = 51 (sd = 16.7) Female: 63%	N = 1338 Depressed: 13%	Administration: Self-report Language: Dutch	DSM-IV CIDI

Abbreviations: C-DIS = Computerised Diagnostic Interview Schedule; CIDI = Composite International Diagnostic Interview; CIS-R = Clinical Interview Schedule (Revised); DISC = Diagnostic Interview Schedule for Children; DSM-III-R = Diagnostic and Statistical Manual (Version III Revised); DSM-IV = Diagnostic and Statistical Manual (Version IV) ; International Classification of Diseases (Version 10); PHQ-2 = Patient Health Questionnaire two-item version; PRIME-MD = Primary Care Evaluation of Mental Disorders; SCAN = Schedule for Clinical Assessments in Neuropsychiatry; SCID = Structured Clinical Interview for DSM

Table 2: Quality assessment of included studies

Study	**Patient selection:** Consecutive or random sample	**Patient selection:** Avoid case-control / avoid artificially inflated base rate	**Patient selection:** Avoided inappropriate exclusions	**Patient selection:** Overall risk of bias	**Index test:** PHQ-2 interpreted blind to reference test	**Index test:** Threshold pre-specified or multiple cut-offs reported	**Index test:** If translated, appropriate translation	**Index test:** If translated, psychometric properties reported	**Index test:** Overall risk of bias
Arroll et al. (2010)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
Chagas et al. (2011)	✓	✓	✓	Low	✓	✓	×	✓	Low
De Lima Osorio et al. (2009)	✓	✓	×	Low	?	✓	?	?	Unclear
	?	?	×	High	?	✓	✓	?	Unclear
De Lima Osorio et al. (2012)	✓	✓	✓	Low	✓	✓	?	?	Unclear
De Man-van Ginkel et al. (2012)									
Delgadillo et al. (2011)	×	✓	✓	Low	✓	✓	n/a	n/a	Low
Fiest et al. (2014)	✓	✓	✓	Low	✓	×	n/a	n/a	High
Inagaki et al. (2013)	×	×	✓	High	?	✓	?	?	Unclear
Kroenke et al. (2003)	×	✓	×	High	✓	✓	n/a	n/a	Low
Liu et al. (2011)	?	✓	?	Unclear	✓	✓	✓	✓	Low
Lowe et al. (2005)	×	✓	✓	Low	✓	✓	✓	✓	Low
Margrove et al. (2011)	×	×	✓	High	✓	✓	n/a	n/a	Low
Phelan et al. (2010)	×	✓	✓	Low	?	✓	n/a	n/a	Unclear
Richardson et al. (2010)	×	×	✓	High	✓	✓	n/a	n/a	Low
Richardson et al. (2010)	×	✓	✓	Low	✓	✓	n/a	n/a	Low
Smith et al. (2010)	?	✓	?	Unclear	✓	✓	n/a	n/a	Low
Thombs et al. (2008)	×	✓	?	Unclear	?	✓	n/a	n/a	Unclear
Tsai et al. (2014)	?	×	✓	High	✓	✓	?	?	Unclear
Williams et al. (2005)	×	?	✓	Unclear	✓	✓	n/a	n/a	Low
Zhang et al. (2013)	?	✓	✓	Unclear	✓	✓	✓	?	Unclear
Zuithoff et al. (2010)	×	✓	✓	Low	✓	✓	✓	?	Low

✓ = criterion met; × = criterion not met; ? = insufficient information to code whether criterion met; n/a = not applicable

[1]If studies reported multiple cut-off points, ‘threshold pre-specified’ is coded as not applicable.

Table 2: Quality assessment of included studies (continued)

Study	**Reference test:** Reference test correctly classifies target condition	**Reference test:** Reference test interpreted blind to PHQ-2	**Reference test:** If translated, appropriate translation	**Reference test:** If translated, psychometric properties reported	**Reference test:** Overall risk of bias	**Flow / timing:** Interval of two weeks or less	**Flow / timing:** All participants receive same reference test	**Flow / timing:** All participants included in analysis?	**Flow / timing:** Overall risk of bias
Arroll et al. (2010)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Chagas et al. (2011)	✓	?	×	✓	Unclear	✓	✓	×	Low
De Lima Osorio et al. (2009)	✓	?	?	?	Unclear	?	✓	✓	Unclear
De Lima Osorio et al. (2012)	✓	?	?	?	Unclear	✓	✓	×	High
De Man-van Ginkel et al. (2012)	✓	✓	?	?	Unclear	✓	✓	×	High
Delgadillo et al. (2011)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
Fiest et al. (2014)	✓	✓	n/a	n/a	Low	✓	✓	×	High
Inagaki et al. (2013)	✓	?	✓	?	Unclear	✓	✓	×	High
Kroenke et al. (2003)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Liu et al. (2011)	✓	✓	?	✓	Low	✓	✓	×	Low
Lowe et al. (2005)	✓	✓	?	?	Unclear	✓	✓	✓	Low
Margrove et al. (2011)	✓	?	n/a	n/a	Unclear	?	✓	×	Unclear
Phelan et al. (2010)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Richardson et al. (2010) [22]	✓	×	n/a	n/a	High	✓	✓	✓	Low
Richardson et al. (2010) [23]	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
Smith et al. (2010)	✓	?	n/a	n/a	Unclear	×	✓	✓	Low
Thombs et al. (2008)	?	✓	n/a	n/a	Unclear	✓	✓	✓	Low
Tsai et al. (2014)	✓	✓	✓	✓	Low	?	✓	×	High
Williams et al. (2005)	✓	×	n/a	n/a	High	✓	✓	✓	Low
Zhang et al. (2013)	✓	✓	?	?	Unclear	✓	✓	×	High

Zuithoff et al. (2010)	✓	✓	✓	✓	Low	?	✓	✓	Low

1 ✓ = criterion met; × = criterion not met; ? = insufficient information to code whether criterion met; n/a = not applicable

Table 2: Quality assessment of included studies (continued)

Study	Patient selection: Applicability	Index test: Applicability	Reference test: Applicability
Arroll et al. (2010)	✓	✓	✓
Chagas et al. (2011)	✓	✓	✓
De Lima Osorio et al. (2009)	✓	✓	✓
De Lima Osorio et al. (2012)	✓	✓	✓
De Man-van Ginkel et al. (2012)	✓	✓	✓
Delgadillo et al. (2011)	✓	✓	✓
Inagaki et al. (2013)	✓	✓	✓
Fiest et al. (2014)	✓	✓	✓
Kroenke et al. (2003)	✓	✓	✓
Liu et al. (2011)	✓	✓	✓
Lowe et al. (2005)	✓	✓	✓
Margrove et al. (2011)	✓	✓	✓
Phelan et al. (2010)	✓	✓	✓
Richardson et al. (2010)	✓	✓	✓
Richardson et al. (2010)	✓	✓	✓
Smith et al. (2010)	✓	✓	✓
Thombs et al. (2008)	✓	✓	✓
Tsai et al. (2014)	✓	✓	✓
Williams et al. (2005)	✓	✓	✓
Zhang et al. (2013)	✓	✓	✓
Zuithoff et al. (2010)	✓	✓	×

✓ = criterion met; × = criterion not met; ? = insufficient information to code whether criterion met; n/a = not applicable

Table 3: Diagnostic test accuracy of the PHQ-2 at cut off point ≥3

	Sensitivity (95% CI)	Specificity (95% CI)	+ve LR (95% CI)	-ve LR (95% CI)	DOR (95% CI)
Arroll et al. (2010)	0.61 (0.53-0.69)	0.92 (0.91-0.93)	7.68 (6.41-9.2)	0.42 (0.35-0.51)	18.3 (12.9-25.8)
Chagas et al. (2011)	0.75 (0.55-0.89)	0.89 (0.80-0.95)	6.83 (3.56-13.1)	0.28 (0.15-0.54)	24.3 (8.22-72)
De Lima Osorio et al. (2009)	0.97 (0.89-1)	0.88 (0.81-0.93)	8.08 (4.93-13.2)	0.04 (0.01-0.14)	213 (50.9-*)
De Lima Osorio et al. (2012)	1 (0.15-1)	0.75 (0.65 – 0.83)	4.08 (2.88 – 5.78)	0 (* - *)	* (1.53 - *)
Delgadillo et al. (2011)	0.68 (0.55-0.79)	0.68 (0.51-0.81)	2.1 (1.3-3.4)	0.47 (0.31-0.72)	4.47 (1.93-10.3)
Inagaki et al. (2013)	0.78 (0.61 – 0.90)	0.85 (0.87 – 0.99)	17.50 (5.72 – 53.6)	0.22 (0.12 – 0.41)	77.3 (19.9– 294)
Kroenke et al. (2003)	0.83 (0.68-0.93)	0.90 (0.87-0.92)	8.28 (6.2-11)	0.19 (0.1-0.37)	43.6 (18.8-101)
Liu et al. (2011)	0.64 (0.49-0.77)	0.94 (0.92-0.95)	9.98 (7.51-13.3)	0.39 (0.27-0.56)	26 (14.1-47.6)
Lowe et al. (2005)	0.87 (0.77-0.94)	0.78 (0.74-0.82)	3.96 (3.26-4.81)	0.16 (0.09-0.3)	24.4 (11.8-50)
Margrove et al. (2011)	0.8 (0.59-0.93)	1 (0.87-1)	* (*-*)	0.2 (0.91-0.44)	* (23.6-*)
Phelan et al. (2010)	0.63 (0.24-0.92)	0.85 (0.74-0.93)	4.24 (1.89-9.5)	0.44 (0.18-1.08)	9.63 (2.12-43.5)
Richardson et al. (2010) [22]	0.74 (0.67-0.79)	0.75 (0.69-0.81)	2.97 (2.31-3.82)	0.35 (0.28-0.44)	8.46 (5.51-13)
Richardson et al. (2010) [23]	0.80 (0.71-0.88)	0.78 (0.73-0.83)	3.63 (2.85-4.62)	0.25 (0.17-0.38)	14.3 (8.13-25)
Smith et al. (2010)	0.77 (0.46-0.95)	0.59 (0.52-0.66)	1.88 (1.33-2.64)	0.39 (0.14-1.06)	4.8 (1.37-16.6)
Thombs et al. (2008)	0.39 (0.32-0.46)	0.93 (0.91-0.95)	5.55 (4.1-7.5)	0.66 (0.59-0.73)	8.4 (0.58-12.3)
Tsai et al (2014)	0.94 (0.72 – 0.99)	0.82 (0.75 – 0.88)	5.34 (3.7 – 7.7)	0.06 (0.01 – 0.45)	79.1 (12.7 - *)

Williams et al. (2005)	0.83 (0.75-0.90)	0.84 (0.78-0.89)	5.13 (3.73-7.06)	0.20 (0.13-0.31)	25.3 (13.6-47.1)
Zhang et al. (2013)	0.79 (0.69 – 0.87)	0.96 (0.94 – 0.97)	19.9 (14.2 – 28.1)	0.21 (0.13- 0.32)	94.6 (50.5 – 177)
Zuithoff et al. (2010)	0.42 (0.34-0.50)	0.94 (0.92-0.95)	6.98 (5.24-9.29)	0.62 (0.54-0.7)	11.3 (7.71-16.6)

1 Note: * Value could not be estimated
2 Abbreviations: -ve LR: Negative likelihood ratio; +ve LR: Positive likelihood ratio; DOR: Diagnostic odds ratio

Table 4: Diagnostic test accuracy of the PHQ-2 at cut off point ≥2

	Sensitivity (95% CI)	Specificity (95% CI)	+ve LR (95% CI)	-ve LR (95% CI)	DOR (95% CI)
Arroll et al. (2010)	0.86 (0.80-0.91)	0.78 (0.77-0.80)	3.95 (3.58-4.35)	0.18 (0.12-0.26)	21.9 (14.0-34.3)
Chagas et al. (2011)	0.93 (0.77-0.99)	0.70 (0.58-0.79)	3.05 (2.16-4.29)	0.10 (0.03-0.39)	29.6 (7.15-*)
De Lima Osorio et al. (2009)	1 (0.94-1)	0.78 (0.70-0.86)	4.64 (3.28-6.57)	0 (*-*)	* (55.6-*)
De Lima Osorio et al. (2012)	1 (0.15 – 1)	0.50 (0.39 – 0.60)	2 (1.64-2.44)	0 (*-*)	* (50.3-*)
De Man-van Ginkel et al (2012)	0.75 (0.50 – 0.91)	0.76 (0.67 – 0.82)	3.09 (2.1 – 4.53)	0.33 (0.15 – 0.71)	9.34 (3.27 – 26.50)
Fiest et al. (2014)	0.40 (0.22 – 0.61)	0.88 (0.82 – 0.92)	3.47 (1.89 – 6.37)	0.67 (0.48 – 0.92)	5.17 (2.15 – 12.50)
Inagaki et al. (2013)	0.78 (0.61-0.90)	0.89 (0.79 – 0.95)	7.50 (3.65 – 15.4)	0.24 (0.13 – 0.44)	31.1 (10.4 – 92.7)
Kroenke et al. (2003)	0.93 (0.80-0.99)	0.74 (0.70-0.77)	3.52 (2.98-4.15)	0.10 (0.03-0.30)	35.4 (11.4-110)
Liu et al. (2011)	0.88 (0.76-0.96)	0.82 (0.80-0.84)	4.87 (4.19-5.65)	0.15 (0.07-0.31)	33.3 (14.3-76.8)
Lowe et al. (2005)	1 (0.95-1)	0.51 (0.46-0.56)	2.04 (1.86-2.24)	0 (*-*)	* (19.2-*)
Phelan et al. (2010)	0.75 (0.35-0.97)	0.67 (0.54-0.79)	2.29 (1.34-3.92)	0.37 (0.11-1.25)	6.15 (1.28-*)
Richardson et al. (2010)	0.90 (0.85-0.93)	0.57 (0.50-0.64)	2.08 (1.77-2.45)	0.18 (0.12-0.29)	11.5 (6.98-18.8)
Richardson et al. (2010)	0.95 (88.8-0.98)	0.58 (0.52-0.64)	2.26 (1.96-2.62)	0.9 (0.04-0.20)	26.5 (10.7-65.2)
Thombs et al. (2008)	0.82 (0.77-0.87)	0.79 (0.76-0.82)	3.91 (3.37-4.53)	0.23 (0.17-0.3)	17.3 (11.8-25.3)
Tsai et al. (2014)	1 (0.81 – 1)	0.49 (0.41 – 0.58)	1.99 (1.69 – 2.33)	0 (*-*)	* (4.55 - *)
Zhang et al. (2013)	0.96 (0.89 – 0.99)	0.57 (0.53 – 0.60)	2.24 (2.06 – 2.44)	0.06 (0.02 – 0.19)	35.8 (11.9 – 108)

Zuithoff et al. (2010)	0.81 (0.75-0.87)	0.76 (0.73-0.78)	3.38 (2.99-3.83)	0.25 (0.18-0.34)	13.7 (9.2-20.5)

Note: * Value could not be estimated

Abbreviations: -ve LR: Negative likelihood ratio; +ve LR: Positive likelihood ratio; DOR: Diagnostic odds ratio

Figure 1: PRISMA Flow diagram outlining study selection

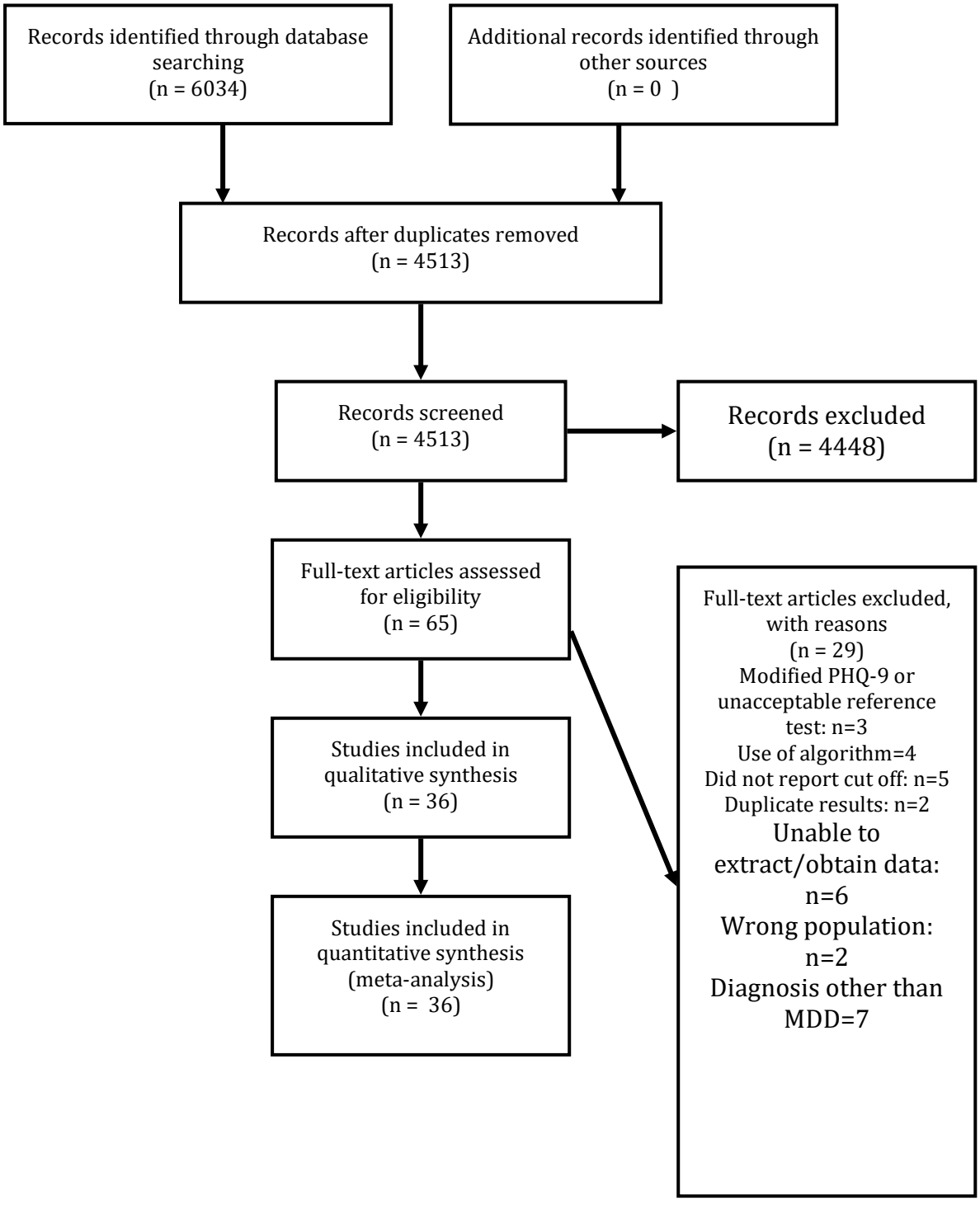


Figure 2. PHQ-2 at ≥3 summary ROC plot of diagnosis of major depressive disorder. Pooled sensitivity and specificity using a bi-variate meta-analysis.

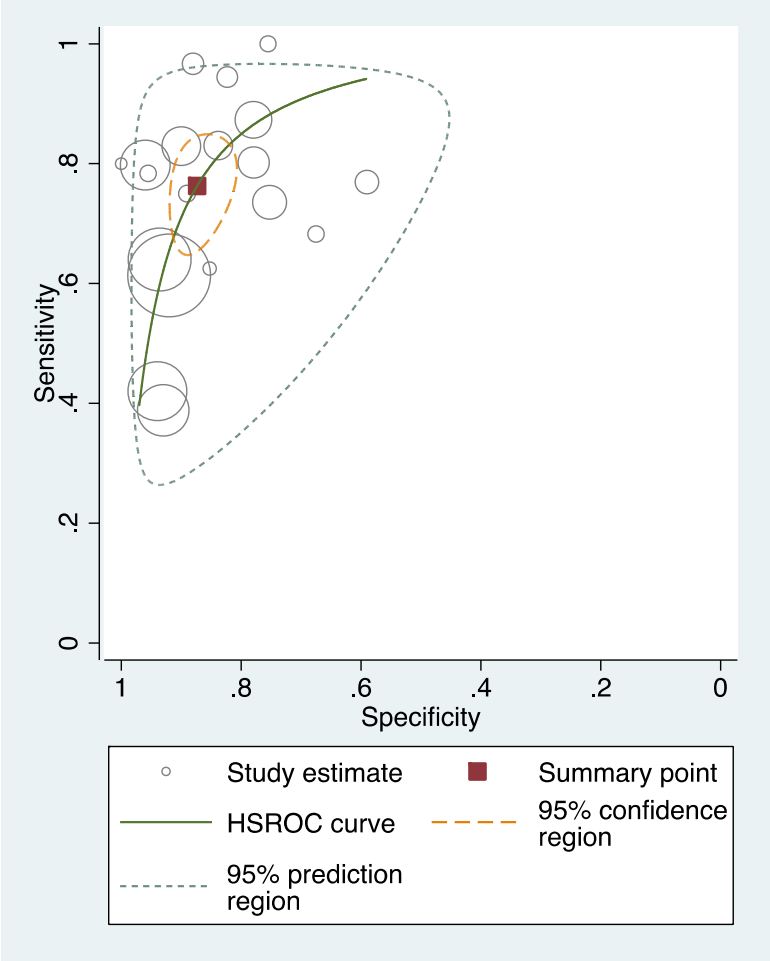


Figure 3. PHQ-2 at ≥2 summary ROC plot of diagnosis of major depressive disorder. Pooled sensitivity and specificity using a bi-variate meta-analysis.

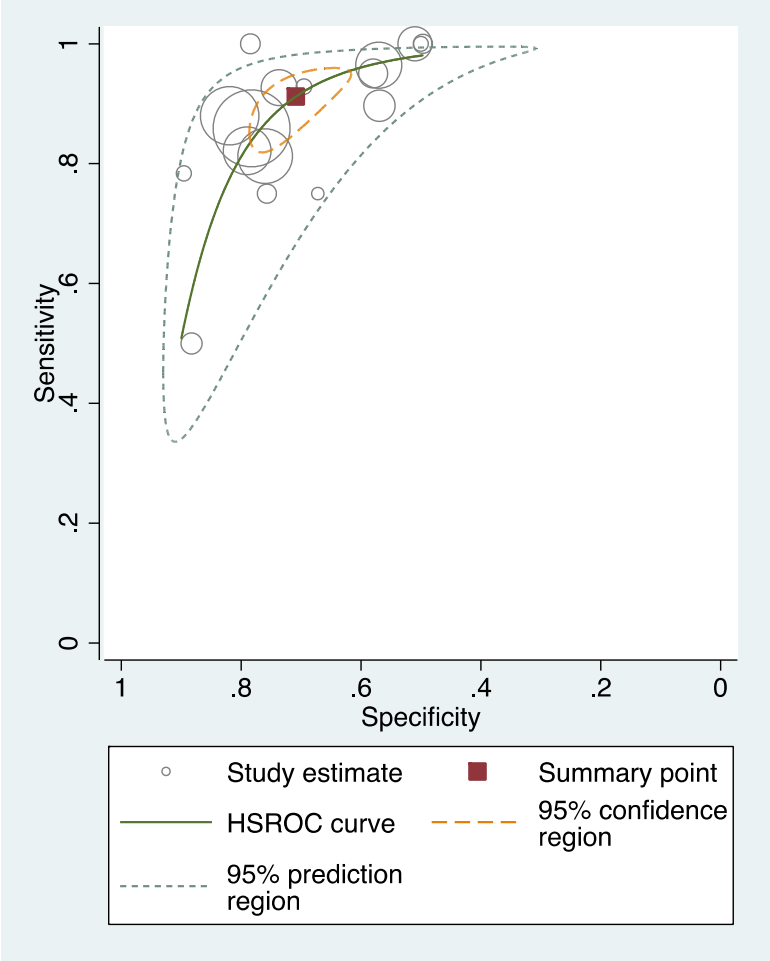


Figure 4: Performance of PHQ-2 at ≥2 using pooled sensitivity and specificity at different prevalence estimates

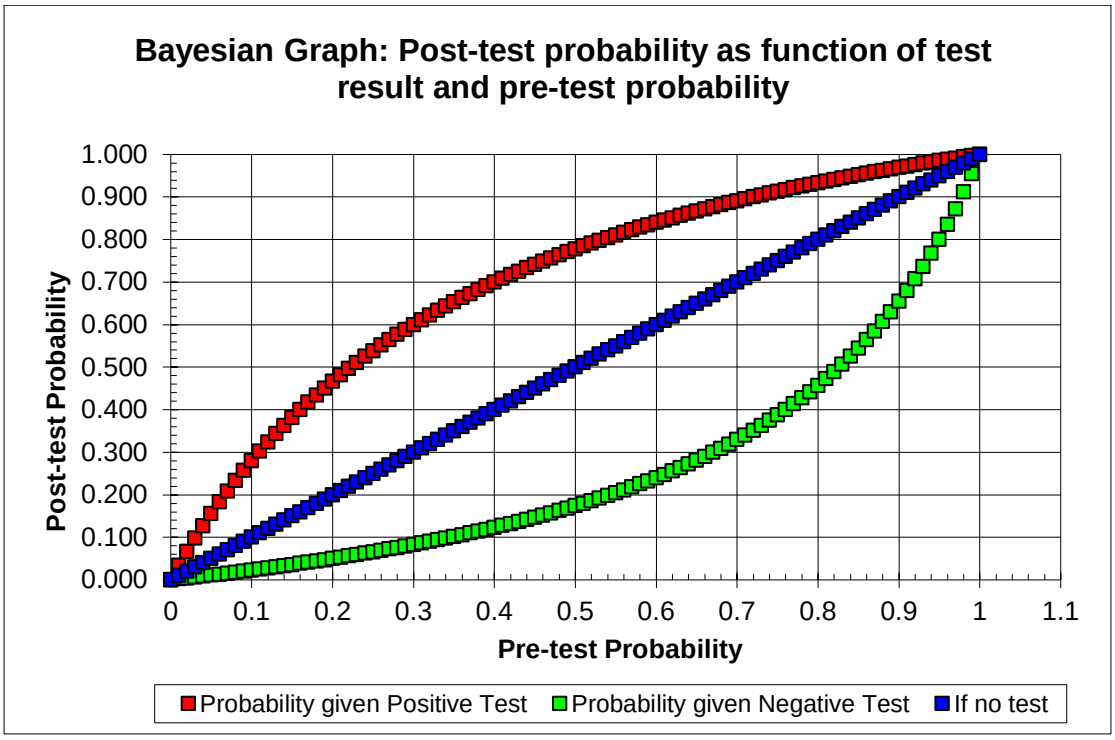
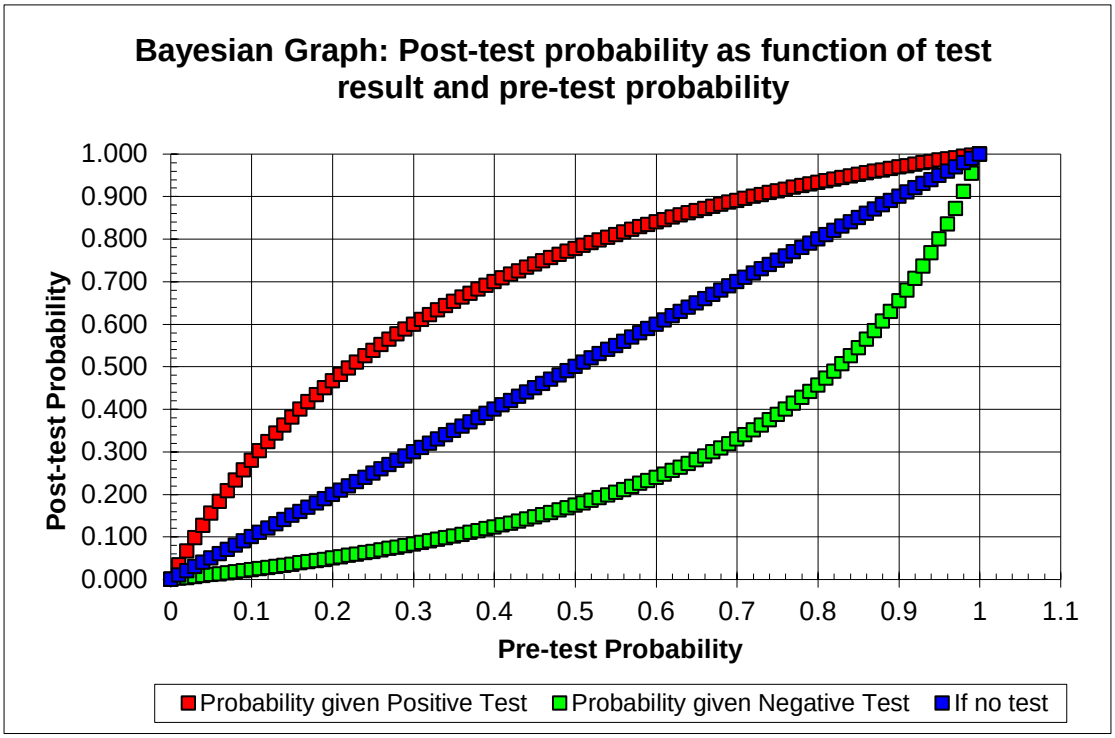


Figure 5: Performance of PHQ-2 at ≥2 using pooled sensitivity and specificity at different prevalence estimates in primary care studies



(Gilbody, Richards, Brealey, & Hewitt, 2007)

Appendix 1: Search terms used in Embase, MEDLINE and PsycINFO

(phq adj5 “2”).ti,ab.

(phq adj5 abbreviate$).ti,ab.

(phq adj5 brief).ti,ab.

(phq adj5 item$).ti,ab.

(phq adj5 short$).ti,ab.

(phq adj5 two).ti,ab.

(patient health questionnaire adj5 “2”).ti,ab.

(patient health questionnaire adj5 abbreviate$).ti,ab.

(patient health questionnaire adj5 brief).ti,ab.

(patient health questionnaire adj5 item$).ti,ab.

(patient health questionnaire adj5 short$).ti,ab.

(patient health questionnaire adj5 two).ti,ab.

(prime md adj5 “2”).ti,ab.

(prime md adj5 abbreviate$).ti,ab.

(prime md adj5 brief).ti,ab.

(prime md adj5 item$).ti,ab.

(prime md adj5 short$).ti,ab.

(prime md adj5 two).ti,ab.

Appendix 2: Excluded studies and reasons for exclusion

Study	Reason for exclusion	Further information
Allgaier et al. (2012)	Reference standard not solely major depression	
Baker-Glenn et al. (2011)	Non-standard PHQ-2 scoring	If either of the two questions were scored as positive, the test was considered positive.
Boyle et al. (2011)	Overlap in sample	Overlap with Richardson et al. (2010)
Brody et al. (1998)	Not PHQ-2	From description of the measure, it is not clear that it is the PHQ-2
Bunevicius et al (2013)	Inadequate reference standard	
Celano et al. (2013)	Inadequate reference standard	
Chen et al. (2010)	Insufficient information to calculate 2*2 table	Sensitivity and specificity reported, but other information needed to calculate 2*2 table such as base rate of depression according to gold standard not reported
De Man-van Ginkel et al. (2012)	Inadequate reference standard	
Elderon et al. (2011)	Overlap in sample	Overlap with Thombs et al. (2008)
Gjerdingen et al. (2009)	Non-standard PHQ-2 scoring	PHQ-2 scored as positive if either question scored ≥2
Hahn et al. (2006)	Not PHQ-2	Uses PHQ-9 not PHQ-2
Hammerton et al. (2014)	PHQ-9/PHQ-2 used to detect recurrent depression	Included patients already known to have depression
Henkel et al. (2003)	Not PHQ-2	Uses PHQ-9 not PHQ-2
Henkel et al. (2004)	Insufficient information to calculate 2*2 table	Sufficient information reported to calculate 2*2 table for ‘any depressive disorder’ but not major depression
Henkel et al. (2004)	Not PHQ-2	Uses PHQ-9 not PHQ-2

Jiang & Hesser (2011)	Inadequate reference standard	PHQ-8 is treated as the reference standard. (In addition, reference standard is 'any depressive disorder' not major depression.)
Kochar et al. (2007)	Not PHQ-2	Uses PHQ-9 not PHQ-2 (In addition, reference standard is clinician diagnosis)
Kroenke & Spitzer (2002)	Overlap in sample	Overlap with Kroenke et al. (2003)
Li et al. (2007)	Not PHQ-2	Although called PHQ-2 it uses different questions to standard PHQ-2 items
Lowe et al. (2005)	Overlap in sample	Overlap with Lowe et al. (2005)
McGuire et al. (2011)	Reference standard not solely major depression	Reference standard diagnosis was either major or minor depression
McManus et al. (2005)	Overlap in sample	Overlap with Thombs et al. (2008)
Mitchell et al. (2009)	Not PHQ-2	Items were from the Structured Clinical Interview for DSM-IV
Mitchell et al. (2008)	Non-standard PHQ-2 scoring	PHQ-2 scored as positive if either question was scored as positive
Mitchell et al. (2010)	Non-standard PHQ-2 scoring	PHQ-2 scored as positive if either question was scored as positive
Monahan et al. (2008)	Inadequate reference standard	PHQ-9 used as the reference standard
Park et al. (2013)	Inadequate reference standard	
Pibernik-Okanovic et al. (2009)	Reference standard not solely major depression	Reference standard diagnosis combines major depression and dysthymia
Richardson et al. (2008)	Overlap in sample	Overlap with Richardson et al. (2010)
Rickels et al. (2009)	Non-standard PHQ-2 scoring	Items are scored yes / no

Rivera-Todaro et al. (2009)	Unable to obtain additional information	Unable to contact authors to obtain further information. Only information available is that contained in the conference abstract.
Robison et al (2002)	Not PHQ-2	Uses the Whooley questions not the PHQ-2
Rollman et al. (2012)	Non-standard PHQ-2 scoring	PHQ-2 scored as positive if either question was scored as positive.
Ryan et al. (2012)	Not PHQ-2	
Smolderen et al. (2011)	Inadequate reference standard	Uses a variety of case records to determine depression status
Tiffin (2010)	Overlap in sample	A review of Richardson et al. (2010)
Wagner et al. (2013)	Insufficient information	Only abstract available
Watson et al. (2009)	Non-standard PHQ-2 scoring	PHQ-2 scored with yes-no response (In addition, reference standard is 'any depressive disorder' not major depression.)