

**Psychometric properties of most commonly used screening and case finding tools for
major depressive disorder in non-psychiatric settings**

Laura Elena Voss (nee Manea)

MD (University of Medicine Cluj-Napoca) MSc (University of Glasgow) MMedSci (Hull
York Medical School) MRCPsych (United Kingdom)

PhD by Publication

The University of Hull and The University of York

Hull York Medical School

March 2021

1 Abstract

Background:

Major depressive disorder (MDD) is the most common mental disorder in medical population, but its identification remains poor. Various screening strategies, involving the use of brief questionnaires to identify patients with unrecognized MDD, have been proposed as a way to improve this. However, the recommended use of specific questionnaires was based on limited or indeed, inaccurate information.

Aim:

To investigate how the most commonly recommended instruments to screen or case find for depression perform in non-psychiatric settings.

Methods:

This thesis is based on six published papers that used rigorous systematic review and meta-analytic research methods to assess the diagnostic accuracy of depression identification instruments recommended by national guidelines.

Main findings:

This work has highlighted significant shortcomings in the existing evidence on depression screening instruments' accuracy. The methodological quality of primary validation studies was generally poor. The included reviews identified that most validation studies have been conducted in samples too small to precisely estimate accuracy and may have selectively published accuracy results from high-performing cut-offs. For the standard cut-off points, the performance of examined instruments was generally poorer than that reported in the original validation studies. Moreover, the diagnostic performance varied by healthcare setting. Presented evidence suggests that the same single threshold might not be appropriate in all settings.

Conclusions:

The research included in this thesis has been used in various national guidelines on identifying depression, and may further change how common depression identification instruments are used. The quality of these reviews has been recognised in subsequently published research. This work has also highlighted important methodological issues that have not been previously identified in diagnostic test accuracy research - the potential allegiance effect in studies co-authored by the original developers of an instrument. Important recommendations are made that are hoped to improve research and clinical practice in this area.

Table of Contents

1	Abstract	2
2	Acknowledgements	5
3	Author’s declaration	6
4	Introduction	10
4.1	Background	12
4.1.1	Depression – a brief synopsis	12
4.1.2	Depression identification - screening, case finding, policies and controversies	13
4.1.3	Screening and case finding instruments for depression	14
4.2	Thesis structure	16
4.3	Aims and objectives	16
5	Research Methods	18
5.1	Search strategy.....	18
5.2	Quality assessment and examination of bias	19
5.3	Data extraction	19
5.4	Summary statistics used to evaluate identification instruments	20
5.5	Data synthesis and statistical analysis	22
5.6	Heterogeneity	25
5.7	Publication bias	26
6	Paper 1	27
6.1	How does this study contribute to the body of knowledge?	27
6.2	Implications of research findings for policy and practice.....	28
6.3	Implications for future research	29
7	Paper 2	31

7.1	How does this study contribute to the body of knowledge?	31
7.2	Implications of research findings to policy and practice	32
7.3	Implications for future research	33
8	Paper 3	35
8.1	How does this study contribute to the body of knowledge?	35
8.2	Implications of research findings to policy and practice	36
8.3	Implications for future research	37
9	Paper 4	38
9.1	How does this study contribute to the body of knowledge?	38
9.2	Implications of research findings for policy and practice.....	39
9.3	Implications for future research	40
10	Paper 5	41
10.1	How does this study contribute to the body of knowledge?.....	41
10.2	Implications of research findings to policy and practice	42
10.3	Implications for future research.....	43
11	Paper 6	44
11.1	How does this study contribute to the body of knowledge?.....	44
11.2	Implications for future research.....	45
12	Final remarks and recommendations	47
12.1	Policy and practice recommendations	48
12.2	Research recommendations	51
13	References.....	56
14	Appendices	70

2 Acknowledgements

I am deeply thankful to my supervisors and mentors Professors Dean McMillan and Simon Gilbody for initiating this endeavor, for their help, support, encouragement, guidance and infinite patience throughout this enterprise and my medical specialty training.

I am also extremely grateful to Dr Stella Morris and Professor Ivana Markova for their invaluable support and encouragement, and for being such inspirational role models.

I would like to thank my co-authors for making this collection of papers possible and from whom I have learned so much.

Finally, I would like to thank the NIHR for funding this work through my NIHR clinical lectureship and to the Humber Teaching NHS Foundation Trust for funding the final writing up stages of this thesis.

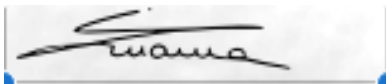
3 Author's declaration

I confirm that this work is original and that if any passage(s) or diagram(s) have been copied from academic papers, books, the internet or any other sources these are clearly identified by the use of quotation marks and the reference(s) is fully cited. I certify that, other than where indicated, this is my own work and does not breach the regulations of HYMS, the University of Hull or the University of York regarding plagiarism or academic conduct in examinations. I have read the HYMS Code of Practice on Academic Misconduct, and state that this piece of work is my own and does not contain any unacknowledged work from any other sources.

This thesis is based on a collection of six papers listed below. While all the papers are co-authored with others, I made significant individual contribution to each of them. Details of these contributions are provided below, signed by myself and the corresponding author, or another major contributory co-author for each paper (for papers where I am the corresponding author).

1. Manea, L, Gilbody, S & McMillan, D 2015, 'A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression' General Hospital Psychiatry, vol 37, no. 1, pp. 67-75.

Contribution of the candidate: I was involved in the conception of the research question and study design; and conducted the literature search. I conducted study selection and data extraction. I also conducted the data analysis, synthesis and interpretation, and prepared the manuscript including subsequent revisions, and approved the final version.



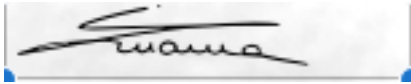
Laura Voss (nee Manea) corresponding author



Dean McMillan

2. Moriarty, AS, Gilbody, SM, McMillan, D & Manea, LE 2015, 'Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis' *General Hospital Psychiatry*, vol 37, no. 6, pp. 567-576.

Contribution of the candidate: I was involved in the conception of the research question and study design; and was involved in conducting the literature search. I was also involved in study selection and data extraction. I conducted the analysis, synthesis, and interpretation of data and prepared the manuscript including subsequent revisions, and approved the final version.



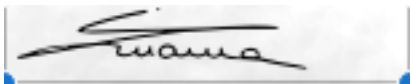
Laura Voss (nee Manea) corresponding author

Dean McMillan

3. Manea, L, Gilbody, S, Hewitt, C, North, A, Plummer, F, Richardson, R, Thombs, B, Williams, B, McMillan, D 2016 'Identifying depression with the PHQ-2: A diagnostic meta-analysis.' *Journal of Affective Disorders* 203, pp. 382-395.

Contribution of the candidate: I was involved in the conception of the research question and study design; and was involved in conducting the literature search. I was also involved in study selection and data extraction.

I conducted the analysis, synthesis, and interpretation of data and prepared the manuscript including subsequent revisions, and approved the final version.

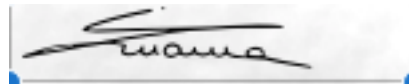


Laura Voss (nee Manea)

Dean McMillan (corresponding author)

4. Bosanquet, K, Bailey, D, Gilbody, SM, Harden, M, Manea, LE, Nutbrown, SE & McMillan, D 2015, 'Diagnostic accuracy of the Whooley questions for the identification of depression: a diagnostic meta-analysis' *BMJ open* 5, pp. 1-12.

Contribution of the candidate: I was involved in all stages of the project from the conception of the research question, study design, writing of the protocol, through study selection, data extraction, quality assessment. I conducted the analysis and data synthesis; and I was involved in preparation of the manuscript including subsequent revisions and approving the final version.

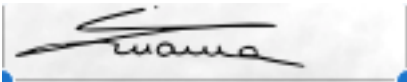


Katharine Bosanquet (corresponding author)

Laura Voss (nee Manea)

5. Pocklington, C, McMillan, D, Gilbody, SM & Manea, LE 2015, 'The Diagnostic Accuracy of Brief and Ultra-brief Versions of the Geriatric Depression Scale: a Meta-analysis' vol 30, no. Supp 1, pp. 1437.

Contribution of the candidate: I was involved in the conception of the research question and study design. I was also involved in study selection and data extraction, analysis, synthesis, and interpretation of data and preparation of the manuscript including subsequent revisions, and approving the final version.

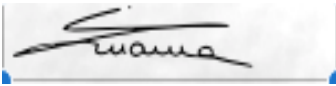


Laura Voss (nee Manea)

Dean McMillan (corresponding author)

6. Manea, LE, Boehnke, JR, Gilbody, SM, Moriarty, AS, & McMillan, D 2017, 'Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis' *BMJ open* 7, pp. 1-23

Contribution of the candidate: I conceived the research question and designed the study. I analysed, synthesised, and interpreted the data. I was also responsible for preparation of the manuscript including subsequent revisions and approved the final version.



Laura Voss (nee Manea) corresponding author

Dean McMillan

I have not submitted any of this work for examination at this or any other institution for another award.

4 Introduction

Up to 10%–20% of patients in medical settings may have a comorbid major depressive disorder (MDD). Research has increasingly shown that unidentified and inadequately treated depression has a major impact on overall health and is a strong indicator of poor prognosis, beyond other health risk factors. (1)

The majority of people with depression are managed in primary care and half of cases are missed (2). In secondary care the prevalence of depression is even higher, but staff typically have less specific training in recognising or managing depression than primary care providers.(3) Thus, improving recognition and management of depression in non-psychiatric settings depression care remains a priority.

Routine depression screening, which involves the use of self-report questionnaires to identify patients with unrecognized MDD who have not been identified as at risk for depression, has been proposed to improve depression identification and management. (4) However, recommendations for routine screening have been made without reference to empirical data demonstrating that it would be clinical and cost-effective. (5) The recommendations and national guidelines have, not surprisingly, been inconsistent, and more recently increasing concerns have been raised that the true diagnostic accuracy of commonly used depression screening tools is poorly understood; therefore, screening may not result in any improvement in patient care.(6)

Routine screening was an important feature of the “detect–treat–improve” paradigm for addressing undetected depression in primary care in the mid-1990s (5) and was subsequently extended to specialty medical settings. This drive led to the development in the mid-1990s of many screening instruments, including the Primary Care Evaluation of Mental Disorders (PRIME-MD), a diagnostic tool copyrighted by Pfizer Inc. (7). The Patient Health Questionnaire (8) was derived from the PRIME-MD and has become the

most commonly used screening measure world wide (including the UK), having been validated in many countries and medical settings. An abundance of validation studies of various depression screening tools in the following years has attracted increasing scrutiny of their quality (9)

This thesis is based on six papers that examine the accuracy of the most commonly used questionnaires in identifying MDD in non-psychiatric settings and the quality of evidence on which they are based. All six reviews and meta-analyses used state of the art systematic review and meta-analytic methods, rigorous quality assessments of the primary studies, and highlighted important methodological issues that have impacted on the reported performance of the instruments with significant implications for clinical practice, guidelines and future research. The full texts of the papers are provided in the Appendix. This integrative chapter clarifies how these papers form a coherent body of work and make an original and significant contribution to knowledge and understanding in this area.

The work presented here has already had a substantial impact in the research community as well as clinical guidelines. Some of the work is now informing national guidelines on identifying and managing depression. For instance, the Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder and the US Preventive Services Task Force for Depression now recommend that the Whooley questions are used to screen in primary and secondary care settings 'in individuals with risk factors' based on the reported psychometric properties and recommendations made in Paper 4 (10) The quality of these reviews has been recognised in subsequently published research (11,12) and many of the recommendations made in these reviews have been explored by well-established international researchers, further widening the knowledge in the area of diagnostic test accuracy assessment. (13–15)

This introduction section comprises a number of subsections.

- The background section includes a synopsis of depression and current screening recommendations, policies and controversies, including uncertainties about the accuracy of instruments currently endorsed by NICE.
- The aims
- Methodology
- A description of the thesis structure.

4.1 Background

4.1.1 Depression – a brief synopsis

Major depression is a commonly occurring and recurrent disorder with significant consequences for the individual and the wider society. (16) It is the most common mental disorder in primary health care and medical specialty population. (17) A large body of evidence has revealed the substantial economic costs of depression. (18–20)

The diagnosis of depression is based on the patient's self-reported experiences, behavior observations reported by relatives or friends, and a mental state examination. The most widely used criteria for diagnosing depression are found in the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM) and the World Health Organization's International Statistical Classification of Diseases and Related Health Problems (ICD). The relevant definitions and classifications according to the two classificatory systems are presented in Appendix 1.

Depression can present with psychological symptoms (anhedonia, feelings of guilt, worthlessness, helplessness and hopelessness, low self-esteem and confidence, suicidality), behavioural and physical symptoms (tearfulness, irritability, social withdrawal, fatigue, anxiety, sleep and appetite changes) (21) and cognitive changes

(poor concentration and attention, pessimistic thoughts, mental slowing and rumination).
(22)

Currently available treatments for depression are effective. (23) In the UK the National Institute for Health and Care Excellence (NICE) produces evidence-based guidance on depression treatments and management, their quality, efficiency and cost-effectiveness. (24) However, given that around 50% of people with depression never consult a doctor, 95% never enter secondary mental health services, and many more people are never recognized as being depressed or treated, there is significant room for improving detection and management of depression in healthcare settings. Unfortunately there is substantial decision uncertainty about the value of screening or case finding for depression in non-psychiatric settings, particularly in primary care. (25)

Screening refers to a strategy to identify people with a particular problem in a population currently without signs or symptoms and would involve administering a screening measure to all people presenting in a healthcare setting. Case finding, in contrast, involves applying a screening measure to a subpopulation known to be at increased risk of a particular problem. This may involve the routine use of a case-finding tool for people with a physical health problem that is associated with an increased risk of depression.(26)

4.1.2 Depression identification - screening, case finding, policies and controversies

Research has consistently identified that depression is under-recognised and under-treated in primary and non-psychiatric secondary care settings, and screening was proposed as a possible solution. Those with positive screening results can then be assessed to ascertain whether they have depression and, if appropriate, offered treatment. (1) Before 2002, routine depression screening has never been recommended by major guidelines. In 2002, the United States Preventive Services Task Force (USPSTF) recommended routine depression screening in primary care settings (27–30)and a

similar recommendation was issued in 2005 by the Canadian Task Force on Preventive Health Care (CTFPHC) (31). In 2010, the UK National Screening Committee (NSC) found that there was no evidence that depression screening would reduce the number of patients with depression or improve depression symptoms, (32) and the UK NICE recommended against routine depression screening (25).

For such a highly prevalent condition that can have substantial implications for individuals, their families, the economy and society, there is remarkable uncertainty for decision-makers about whether to screen or case find for depression. There is still substantial disagreement between different national guidance about the benefits of these strategies. US guidelines recommend a form of screening, offered to all regardless of level of risk, if there are appropriate structures and processes in place to manage those identified as depressed. (27) UK NICE guidance, while not recommending this general screening approach, recommend an alternative strategy involving the use of brief case-finding instrument for people deemed at increased risk, such as those with chronic physical health problems (25,26) This is conflicting with the UK NSC conclusion that there is insufficient evidence to recommend the adoption of screening for depression in general, and the lack of robust evidence for case finding among high risk populations. (33) Canadian guidelines (31) strongly caution against the use of any form of screening or case finding for depression, because of, among other concerns, a lack of understanding about the potential harms of screening. (34) Potential harms include over-treatment, pathologising normal responses to life events, and the diversion of resources from the management of depression that has already been identified because of its severity or marked impact on functioning. (35) Of course, central to this debate is the diagnostic accuracy of the instruments recommended to screen or case-find for depression.

4.1.3 Screening and case finding instruments for depression

Screening and case-finding instruments are validated psychometric measures that are

used to identify people with previously unrecognised depression. The current NICE guidelines is informed by a review of the diagnostic accuracy of the most commonly used instruments in the UK that has concluded that, based on the available evidence, the Whooley questions, the PHQ-9 and the GDS appeared to perform better. However, an important limitation of the available evidence was the very high heterogeneity found (24)

Given the gaps in available evidence, the reviews included in this thesis examined the diagnostic properties of the instruments which were deemed better performing by current NICE guidance: the PHQ-9 (by far the most commonly used case finding tool in the UK at present), its ultra-brief version the PHQ-2 and another ultra-brief tool recommended by current guidance (the Whooley questions), as well as the commonly used case finding tool in elderly population – the GDS.

The diagnostic accuracy is determined against ‘gold standard’ diagnoses defined as a DSM or ICD diagnosis of depression. Whilst in clinical settings, MDD is most commonly diagnosed through clinical judgment and non-structured interviews, standardized diagnostic interviews, including semi-structured and fully structured interviews are used in research settings. Semi-structured interviews are similar to a guided diagnostic consultation. Standardized questions are asked, but the clinician may ask additional questions and use clinical judgement to decide whether symptoms are present. (36) The Structured Clinical Interview for DSM (SCID) (37) and Schedules for Clinical Assessment in Neuropsychiatry (SCAN) (38) are commonly used semi-structured interviews in research. Fully structured interviews typically involve standardized questions that are read verbatim, without additional questions. (36) They are designed to provide greater standardization, but with less flexibility and do not involve additional clinical judgment, and can, therefore, be administered by lay interviewers. (36) Some commonly used structured interviews in research include the Composite International Diagnostic Interview (CIDI) (39) and the Diagnostic Interview Schedule (DIS).(40) The Mini International Neuropsychiatric Interview (MINI) is also a fully structured interview, but is

designed to be administered in less time and is consequently overinclusive, generating a higher rate of false-positive diagnoses (41,42)

Both fully structured and semi-structured interviews are deemed appropriate reference standards for MDD classification in research.(36) However, as identified in this thesis, it is likely that different interview formats, may lead to different diagnostic patterns and could increase misclassification.

4.2 Thesis structure

This thesis is divided into 9 further separate sections. The next section is a summary of the overarching methodology used in the included reviews. The following 6 sections (6-11) summarise the contribution of each paper to the body of knowledge. Section 12 summarises the identified quality and methodological issues and implications for practice and policy, and draws together final remarks and recommendations.

4.3 Aims and objectives

The overall aim of the research included in this thesis was to examine all the available evidence on diagnostic accuracy of the instruments for depression screening or case finding deemed as most used and best performing by currently published NICE guidelines, and the quality of evidence on which these results are based.

The more specific objectives are:

1. To quantify the diagnostic accuracy of these screening and/or case finding instruments (SCFIs) for MDD
2. To assess the quality of primary validation studies of these instruments, and explore methodological issues that may affect the reported diagnostic accuracy in these studies.

5 Research Methods

The reviews included in this thesis followed the Centre for Reviews and Dissemination (CRD) guidelines for the systematic review of diagnostic studies (43) and the Cochrane Handbook of DTA Reviews (<http://srdta.cochrane.org/handbook-dta-reviews>). (44)

5.1 Search strategy and study selection

We used the searching guidance from the Cochrane Handbook of DTA Reviews (<http://srdta.cochrane.org/handbook-dta-reviews>). The search terms included a list of named SCFIs for depression (e.g., PHQ-2, PHQ-9, Whooley, GDS). We searched a range of databases including MEDLINE, PsycINFO, EMBASE. Grey literature was identified from searches of databases such as Dissertation Abstracts, OAISTER and ZETOC. We also checked the reference lists of included studies to identify any additional studies.

Studies were selected using a pre-piloted form based on the PICO criteria (roughly outlined below and adapted for each review).

Overarching Inclusion/Exclusion Criteria

Population: Adults (including young adults) in a non-psychiatric setting (primary care, general hospital settings, community settings)

Screening test: The evaluated brief depression SCFI

Reference test/gold standard: A standardised diagnostic interview conducted according to internationally recognised criteria, such as the ICD system, versions of the DSM or Research Diagnostic Criteria.

Outcomes: Sufficient data to calculate a 2x2 contingency table for at least one cut-off point on the screening measure against a gold standard

Study design: Cross-sectional, case control, cohort studies and RCTs (where screening measures are used as a method of recruitment)

No restrictions were made in terms of publication status, publication year or language.

All identified citations were first assessed based on title and abstract. When possible, this was done by two reviewers. 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. At each stage any disagreements were resolved by consensus and where necessary arbitration by further reviewers. Where necessary authors were contacted to provide further clarification or to obtain additional information.

5.2 Quality assessment and examination of bias

Quality assessment was performed for each review included in this thesis using the updated tool for Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2), (44) which was designed for evaluating the risk of bias and applicability of primary diagnostic accuracy studies when conducting systematic reviews. It covers the areas of patient selection, index test, reference standard, flow and timing. This tool was slightly adapted for these reviews, and two independent reviewers carried out quality assessments. For each paper, I will discuss in detail the methodological quality of included studies and identify potential sources, of bias as well as their impact on the results.

5.3 Data extraction

Two reviewers extracted all data independently. Any disagreements were resolved by discussion with a third reviewer.

5.4 Summary statistics used to evaluate identification instruments

Sensitivity and specificity

The sensitivity of an instrument refers to the proportion of those with MDD who test positive. Specificity refers to the proportion of those who do not have MDD and test negative. Sensitivity and specificity do not necessarily depend on prevalence of depression. For example, sensitivity is concerned with the performance of an identification test conditional on a person having depression. Therefore, the higher false positives often associated with samples of low prevalence will not affect such estimates. The advantage of this approach is that sensitivity and specificity can be applied across populations. (45) However, the main disadvantage is that clinicians tend to find such estimates more difficult to interpret. (45)

When describing the sensitivity and specificity of the different instruments, values above 0.9 are considered 'excellent', 0.8 to 0.9 'good', 0.5 to 0.7 'moderate', 0.3 to 0.5 'low', and less than 0.3 'poor'. (46)

Receiver operator characteristic (ROC) curves

The qualities of a particular tool are summarised in a ROC curve, which plots sensitivity (expressed as a percentage) against (100-specificity).

A test with perfect discrimination would have a ROC curve that passed through the top left-hand corner. A perfect test would have an AUC of 1, and a test with AUC above 0.5 is better than chance. These measures are based on sensitivity and 100-specificity, therefore theoretically they are not affected by prevalence.

Negative and positive predictive values (PV)

Negative predictive value is the proportion of people scoring negative on the test who will not have the disease. Positive predictive value is the proportion of people scoring positive on the test who will have the disease. For both positive and negative predictive values, prevalence explicitly forms part of their calculation. When the prevalence of a disorder is low in a population this is generally associated with a higher negative predictive value and a lower positive predictive value. (47) Therefore, although these statistics are concerned with issues probably more directly applicable to clinical practice (for example, the probability that a person with a positive test result actually has depression), they are largely dependent on the characteristics of the population sampled and cannot be universally applied. (47)

Negative and positive likelihood ratios (LR)

Negative and positive likelihood ratios are also not dependent on prevalence. LR- is calculated by sensitivity/1-specificity and LR+ is 1-sensitivity/ specificity. A value of LR+ higher than 5 and LR- lower than 0.3 suggests the test is relatively accurate.(48)

Diagnostic odds ratios (DOR)

The DOR is LR+/LR-; a value of 20 or greater suggests a good level of accuracy. (48)

Youden's index

Youden's index is a measure for diagnostic accuracy and a global measure of a test performance. Youden's index is calculated by deducting 1 from the sum of test's sensitivity and specificity expressed not as percentage but as a part of a whole

number: $(\text{sensitivity} + \text{specificity}) - 1$. For a test with poor diagnostic accuracy, Youden's index equals 0, and in a perfect test Youden's index equals 1. Youden's index is not sensitive for differences in the sensitivity and specificity of the test, which is its main disadvantage.

Youden's index is not affected by the disease prevalence, but it is affected by the spectrum of the disease, as are also sensitivity specificity, likelihood ratios and DOR. (49)

5.5 Data synthesis and statistical analysis

For each screening tool, the range in sensitivity, specificity and likelihood ratios were calculated, together with possible ranges in positive and negative predictive values which were calculated based on a number of different estimates of disease prevalence and varying cut-off points.

2x2 tables were constructed for each scoring method or cut-off point reported by studies and true positive, true negative, false positive and false negative results were computed.

Meta analytical summaries

The aims of a meta-analysis are to compute and compare estimates of the expected diagnostic accuracy of a test and investigate the variability of results between studies. A choice needs to be made of which summary statistics are to be computed.

In a systematic review it is likely that the collected data will be at a mix of different positivity thresholds. Often tests are evaluated at different thresholds in different studies. Presentation of results at multiple thresholds within a single study is also

encountered, with some studies presenting estimates of ROC curves which show the accuracy of the test at all possible thresholds. In addition, selective reporting of thresholds identified to optimise test accuracy can introduce bias if they are selected in a data driven manner.(13)

A key principle underlying the choice of statistical summary in meta-analysis of test accuracy is that the sensitivity and specificity of a test will vary as the positivity thresholds varies, graphically presented using a ROC curve. The hierarchical models recommended for meta-analysis for DTA reviews account for correlation between sensitivity and specificity observed across studies which is due to the functional relationship between sensitivity and specificity as the threshold varies within each study. This occurs regardless of whether a summary ROC curve or a summary point is the output of choice.

A review author needs to decide whether to use all the studies available to estimate the curve (in which case the meta-analysis will estimate the summary ROC curve) or to estimate a summary sensitivity and specificity point on this curve at a chosen threshold. Estimating summary sensitivity and specificity by pooling studies which mix thresholds will produce an estimate that relates to some notional unspecified average of the thresholds that occur in the included studies, which is clinically unhelpful and should be avoided.(50)

Therefore, the two main strategies to handle mixed and variable thresholds in an analysis are:

- Estimating summary sensitivity and specificity of the test for a common threshold, or at each of several different common thresholds. Each study can contribute to one or more analyses depending on what thresholds it reports. Studies which do not report at any of the selected thresholds are excluded.

- Estimating the underlying ROC curve which describes how sensitivity and specificity trade-off with each other as thresholds vary. In this case one threshold per study is selected to be included in the analysis.(50)

The choice of analytical approach will be influenced by the variation of thresholds in the available studies. For example, if there is little consistency in the thresholds used, meta-analyses which restrict to common thresholds will contain very little data, and estimating a summary ROC may be preferred. If there is little variation in threshold between studies attempting to fit a summary ROC curve will be difficult as the points are likely to be too tightly clustered in ROC space.

It is reasonable to estimate both SROC curves and average operating points in a review, as they may complement each other in providing clinically useful summaries, and powerful ways of detecting effects. Separate analyses of test data at different thresholds may be used to provide clinically informative estimates of sensitivity and specificity, whereas including all studies to estimate how summary ROC curves depend on covariates or test type will be the most powerful way to test hypotheses and investigate heterogeneity. (50)

We used bivariate diagnostic meta-analysis to obtain pooled estimates of specificity and sensitivity and their associated 95% confidence intervals. The bivariate model is a 2-level model which takes into account the precision by which differences in sensitivity and specificity have been measured while incorporating and estimating the amount of between-study variability in both sensitivity and specificity (random effects model).(51)

Screening test scores are considered continuous, with many different possible cut-off scores. ROC curves are the most informative way of representing the inherent trade-offs between sensitivity and specificity for a test or diagnostic instrument. (52) Summary ROC curves (sROC) (51) were constructed using the bivariate model to produce a 95% confidence ellipse within ROC space. (53) Each data point in the summary ROC space

represents a separate study, unlike a traditional ROC plot that explores the effect of varying thresholds on sensitivity and specificity in a single study.

5.6 Heterogeneity

It is essential to evaluate heterogeneity (clinical and methodological differences between the studies) in a meta-analysis. Heterogeneity of the DOR was investigated using the I^2 statistic and through visual examination of plots of study results. If studies (and cut points within studies) were homogenous in terms of sensitivity and specificity then the pooled sensitivity, specificity and likelihood ratios with associated 95% confidence intervals were calculated. The model was fitted using a generalised linear mixed model approach to the bivariate meta-analysis of sensitivity and specificity.⁽⁵⁴⁾ This approach uses the exact binomial distribution to describe the within-study variability of sensitivity and specificity.⁽⁵¹⁾ The generalised linear mixed model approach we used corresponds to the approach to fitting the Hierarchical Summary Receiver Operating Characteristic (HSROC) model.⁽⁵⁵⁾

If significant heterogeneity was evident across studies and measures then we sought to explore its causes. First, summary sROC curves were constructed and visually inspected to identify studies that laid outside the 95% confidence ellipse. Secondly, further analyses were conducted using D ($\log(DOR)$). A weighted multivariate linear meta-regression analysis was used, with weights proportional to the reciprocal of the variance of D representing the within-study variation. 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.

Variables that were considered in the analysis included: reference test, sample source, and baseline prevalence of depression, as well as quality features including study design,

method of patient selection, method of verification, and interpretation of tests. Important sources of heterogeneity are predictive in a meta-regression analysis, and reduce the level of between-study heterogeneity. Given the large number of analyses carried out, we reported only the variables that were significant, where the P-value was lower than 0.05. All analyses were conducted using Stata, including the user-written Stata commands *metandi* (56) and *metareg* (57).

5.7 Publication bias

Larger studies or studies reporting significant or interesting results are more likely to be published and eventually included in systematic reviews. (58) These problems are known as publication bias. This was examined using Begg funnel plots of log diagnostic odds ratio versus the inverse of variance. (59) A funnel plot compares the effect size against a measure of the studies size and is used to investigate any potential sources of bias within the studies.

6 Paper 1

Manea, L, Gilbody, S & McMillan, D 2015, 'A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression' *General Hospital Psychiatry*, vol 37, no. 1, pp. 67-75

Paper 1 aimed to examine the diagnostic accuracy of the PHQ-9 using diagnostic algorithm (including the comparison of the accuracy of this scoring method with the summed item scoring method at cut-off point 10) and the quality of the primary validation studies. The difference between scoring methods is explained in the paper. The results are presented in section 2 of the paper.

6.1 How does this study contribute to the body of knowledge?

This study makes an original contribution to literature in a number of ways. This was the first systematic review and meta-analysis to summarise the performance of the PHQ-9 using the algorithm scoring method, including a direct comparison of the performance of the PHQ-9 algorithm in primary care versus secondary care. Moreover it is the first review to directly compare performance of the two recommended scoring methods (algorithm and summed item scoring method using the recommended cut-off point of 10). The comparison established that the PHQ-9 diagnostic algorithm is less accurate than that using the standard cut-off of ≥ 10 . This study highlighted that the baseline prevalence of MDD in the study population is a significant factor to consider when interpreting validation results and identifies some important methodological issues (discussed below).

Paper 1 is a highly cited paper. According to the Web of Science (a well-known academic platform that provides access to multiple databases that provide comprehensive citation data for many different academic disciplines) 'as of November/December 2019, this **highly cited paper** received enough citations to place it in the top 1% of its academic field based on a highly cited threshold for the field and publication year.' A cited reference

search by title as of 7 May March 2024 reveals 493 citations. Google scholar identified 693 citations of this paper.

Given the very high number of citations, I will summarise in the following 2 sections how the findings of this paper have been utilised in the most relevant citing peer review papers identified by the Web of Science.

6.2 Implications of research findings for policy and practice

The paper demonstrated, using rigorous systematic review and meta-analytic methods that the algorithm method of scoring the PHQ-9 leads to problematically low sensitivity. In both primary care and hospital settings, pooled sensitivity was around 0.55, which is lower than reported in the initial validation study. In practice this would translate in a high probability that the algorithm method would miss many patients with MDD. Interestingly, the only significant variable that was predictive in the meta-regression analysis was the base rate of depression.

In studies directly comparing the algorithm and the standard cut-off point of ≥ 10 , the latter had a better sensitivity (0.77) and maintained good specificity (0.85), hence providing better diagnostic performance for screening purposes or where a high sensitivity is needed. However, caution is needed in interpreting these results because the level of heterogeneity was substantial.

As highlighted in the introductory chapter, it is paramount that screening or case finding measures improve the management of depression. Researchers and clinicians have used the results of this paper to inform the decision to screen in primary or secondary healthcare settings, and in order to balance potential negative effects of screening. (60,61)

The majority of articles that have cited Paper 1 have used the summary statistics presented in this paper, rather than those reported in the original validation study - see for example (62–92).

Some of these studies have used the results presented in this paper to choose the summed-item scoring method rather than the algorithm in order to screen or case find for depression (63–66,68,71,72,74–76,93–96)

This paper followed previous recommendations to summarize diagnostic properties of the PHQ-9 for different scoring methods using a bivariate meta-analysis. It further highlighted the need to summarise the diagnostic properties of the PHQ-9 for different scoring methods, in different settings or populations, which has been addressed by various researchers since this paper has been published – see for example (97,98).

6.3 Implications for future research

This paper highlights a significant coding issue that developers of brief SCFIs should consider. Paper 1 identified that a possible explanation of the low sensitivity of the algorithm method could lie in the proposed coding strategy, which, with the exception of the suicidal ideation question, determines items scored 2 (more than half the days) as meeting depression criteria, whereas items scored as 1 (several days) do not meet the criteria. Distinguishing between these response categories may be confusing for the respondent and implications are discussed in the paper.

This paper identified a number of methodological issues. A significant issue is the design used by some papers in which participants who were more likely to be depressed were also more likely to receive the reference standard. The lack of detail in the reporting of studies 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 conducted blind to the PHQ-

9.

The growing number of diagnostic test accuracy (DTA) meta-analyses has attracted greater scrutiny from researchers. As well as focusing on ensuring good quality of primary care validation studies, researchers should also ensure good quality of DTA meta-analyses. This paper was cited as one of the few identified systematic reviews and DTA meta-analyses that fulfilled the criteria on the highest number of AMSTAR items (a measurement tool to assess the methodological quality of systematic reviews). (11,12,99) The methodology used in this review has been subsequently used in other studies. (100)

On the other hand, a potential limitation of this paper may be not accounting for different reference standards. (98,101). Influenced by this paper, He et al. conducted an individual patient data (IPD) meta-analysis of the PHQ-9 algorithm validation studies that estimated sensitivity and specificity for the original and modified PHQ-9 diagnostic algorithms for all patients, separately by studies that used semi-structured, structured, and MINI reference standards. Interestingly however, their study identified that sensitivity and specificity estimates were not statistically significantly different for any reference standard category when restricted to participants not currently diagnosed or receiving treatment for depression. (101)

As in our study, the main conclusion of the analysis carried out by He et al. was that the PHQ-9 algorithm sensitivity was low across reference standards and subgroups, although specificity was high and that, overall, the accuracy of the PHQ-9 diagnostic algorithms did not compare favorably to that of the PHQ-9 using the standard cut-off of ≥ 10 . (101)

7 Paper 2

Moriarty, AS, Gilbody, SM, McMillan, D & Manea, LE 2015, 'Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis' *General Hospital Psychiatry*, vol 37, no. 6, pp. 567-576.

Paper 2 examined the diagnostic accuracy of the PHQ-9 using the summed item scoring method at different cut-off points. The results are presented in Section 3 of the paper.

7.1 How does this study contribute to the body of knowledge?

The aims of this paper were to establish the diagnostic performance of the PHQ-9 at the standard cut-off point (10), to compare the diagnostic performance of the PHQ-9 at the standard cut-off point in different clinical settings and to attempt to evaluate the diagnostic performance of the PHQ-9 at other cut-off points. The study was a follow-up of a previous review that has been highly cited since its publication with over 1300 citations in Web of Science and 1844 Google Scholar citations (102). This study was carried out following a significant increase in the number of validation studies since our first meta-analysis was published, and aimed to elucidate some of the issues identified, in particular the anomalous psychometric properties at different cut-off points due to selective reporting.

The results of this paper showed that the sensitivity of the PHQ-9 at cut-off point 10 is lower than that reported in the original validation study, whereas the specificity is similar.

Heterogeneity was found to be consistently high. Subgroup analyses evidenced that the diagnostic properties of the test differed by settings at the standard cut-off point; however, heterogeneity in each of these subgroup analyses remained high.

Our previous diagnostic meta-analysis of the PHQ-9 showed that the same single threshold might not be appropriate in all settings.(102) This paper aimed to explore whether different cut-off points perform differently in different settings. Due to selective reporting of cut-off points no firm conclusions could be drawn. Nonetheless this review (like most included in this thesis) highlighted the magnitude of selective reporting practices in DTA research. This finding, and the recommendation that journals publishing validation studies of screening or case-finding instruments request that data on all cut-off points are provided, and that this is made a condition of publication, are important contributions to knowledge in this area of research.

Paper 2 has been cited in at least 245 peer reviewed journal articles according to Web of Science, as of 7 May 2024, and 354 Google Scholar citations. As with Paper 1, given the high number of citations in the following 2 sections I will provide only a summary of how the findings of this paper have been used in the most relevant citing peer review papers identified by the Web of Science.

7.2 Implications of research findings to policy and practice

This paper showed that the use of the same cut-off point might not be appropriate in all settings. This finding was endorsed by the developer of the PHQ-9 in a study published in 2016. (103) Paper 2 demonstrated that the initially recommended cut-off point of 10 might yield a high rate of false negative results in hospital settings, while in primary care might generate more false positive cases. This may have consequences for the use of the PHQ-9 as a SCFI because, often, high sensitivity is required in such circumstances to ensure that few people with depression are missed. One strategy in such circumstances is to lower the cut-off point to increase sensitivity. However, this could not be recommended because it is unclear how the test performs at alternative cut-off points, because of selective reporting. This widespread practice, highlighted by this paper, has

prompted researchers to use different methodologies (like IPD) in order to explore accuracy at different cut-off points.(13) Moreover, by carrying out an IPD meta-analysis of the PHQ-9 Levis et al. developed a web-based tool (depressionscreening100.com/phq) that can be used to estimate the expected number of positive screens and true and false screening outcomes in primary care, which can be a useful clinical guide in this setting. (104)

A substantial caveat that applies to any recommendations about the performance of the PHQ-9 at the standard cut-off point, however, is that the pooled accuracy estimates were associated with high levels of heterogeneity which could not be fully explored even by subgroup analyses carried out in the much larger IPD sample. (104)

The majority of papers citing this study have chosen to use cut-off point 10 based on the psychometric properties reported in the study (rather than those reported by the initial validation study). The following are a few examples of studies that used this cut-off as a threshold for MDD in various clinical settings, and/or quoted summary characteristics for cut-off point of 10, based on the results of our study: (72,105–121)

7.3 Implications for future research

Our first diagnostic meta-analysis of the PHQ-9 published in 2012 was the first paper to highlight the extent of selective reporting in validation studies of the PHQ-9. (9) This has attracted a lot of attention from researchers in the field, leading to a significant change in the reporting guidelines for systematic reviews and DTA meta-analyses and recommendations to researchers in this area (122) Paper 2 specifically revisited the issue of selective reporting following the publication of many more validation studies and clearly recommended that future validation studies should report the full range of cut-off points. Paper 2 showed that, despite recommendations made in our previously published DTA meta-analyses, selective reporting practices remain widespread; therefore the new

PRISMA-DTA guidelines (122), published following our reviews and subsequent IPD meta-analyses, are hoped to address this significant problem in DTA research (identified by the studies that form this thesis).

Similarly to Paper 1, this paper was cited as one of the few systematic reviews and DTA meta-analyses that fulfilled the criteria on the highest number of AMSTAR items. (11,12,99)

The methodological quality of the studies included in the review was variable. The fact that many studies overselected people who were likely to be depressed, which may have introduced partial verification bias, is particularly concerning. The reported sensitivity and specificity in these studies are likely to be inflated. Based on Paper 2, Thombs et al. showed that exaggeration of the prevalence of depression is disproportionately high in low-prevalence populations and blurs distinctions between high- and low-prevalence populations. (123)

Another significant issue was the quality of translation of studies conducted in non-English speaking countries. Very few studies describe the translation process of the PHQ-9 and/or the gold-standard measure used and whether the translations were validated. Poor translation and lack of translation validation can significantly threaten the validity of an instrument; therefore validated translation procedures should be followed.

8 Paper 3

Manea, L, Gilbody, S, Hewitt, C, North, A, Plummer, F, Richardson, R, Thombs, B, Williams, B, McMillan, D 2016 'Identifying depression with the PHQ-2: A diagnostic meta-analysis.' *Journal of Affective Disorders* 203, pp. 382-395.

Paper 3 examined the diagnostic accuracy of the PHQ-2. The results are presented in section 3 of the paper.

8.1 How does this study contribute to the body of knowledge?

This paper established the diagnostic performance of the PHQ-2 at the standard cut-off point (3), and an alternative cut-off point (2) identified post-hoc to be more suitable for screening purposes (i.e. better sensitivity).

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. This paper suggests that the accuracy of the PHQ-2 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. This, however, was not necessarily linked to higher specificity

Paper 3 has been cited in 144 peer reviewed journal articles according to Web of Science as of 7 May 2024 and 222 Google Scholar Citations. The vast majority of the citing articles have chosen to use cut-off point 2 (rather than 3), based on the psychometric properties reported by this study. The following are a few examples of studies that used this cut-off as a threshold for MDD in various clinical settings, and/or quoted summary characteristics of the PHQ-2 based on the results of our study, rather than those reported by the original validation study: (124–134)

8.2 Implications of research findings to policy and practice

This paper suggests that the PHQ-2 at ≥ 2 , rather than at the recommended cut-off point of ≥ 3 , may have value in ruling out depression. Lowering the cut-off point increases sensitivity. 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 and the resources available to further assess those who screen positive.

Prevalence estimates of the validation studies vary substantially. It is likely that the higher estimates are related to sampling strategies that over-selected people who were likely to be depressed. As prevalence falls, the proportion of people who score positively but are not depressed will increase; therefore, in low prevalence populations (e.g, primary care) this cut-off point may generate high false positive rates and would render the instrument of limited use for opportunistic screening. However, as the prevalence increases, 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 a number of caveats to this conclusion, which are discussed in section 4 of the paper.

8.3 Implications for future research

Similarly to previous papers, this paper highlights the need to improve the quality of primary validation studies. Variations in study quality, however, did not appear to be related to outcome according to the meta-regression for cut-off point ≥ 3 .

The lack of detail in the reporting of studies made it difficult to assess some of the QUADAS-2 criteria, in particular whether the reference standard was conducted blind to the PHQ-2. 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.

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. Paper 2 (discussed in the previous chapter) has identified 36 validation studies and most of these do not specifically report the psychometric properties of the PHQ-2. This issue was specifically addressed by an IPD meta-analysis published in 2020 that was able to establish that for PHQ-2 scores of 2 or greater followed by PHQ-9 scores of 10 or greater sensitivity was not significantly different to PHQ-9 scores of 10 or greater alone, and specificity (0.87) was significantly better; therefore, in circumstances where screening procedures allow for quick calculation of PHQ-2 scores before presenting remaining PHQ-9 items (e.g. electronic administration), the combination may be a resource-efficient approach.⁽¹³⁵⁾ This IPD meta-analysis (of 48 studies) has also reported diagnostic properties for the PHQ-2 remarkably similar to those reported in this paper.

9 Paper 4

Bosanquet, K, Bailey, D, Gilbody, SM, Harden, M, Manea, LE, Nutbrown, SE & McMillan, D 2015, 'Diagnostic accuracy of the Whooley questions for the identification of depression: a diagnostic meta-analysis' *BMJ open*, pp. 1-12.

Paper 4 examined the diagnostic performance of the original two-item Whooley questions and their combination with an additional help question. The results section of the paper presents the main findings.

9.1 How does this study contribute to the body of knowledge?

This paper is the first systematic review and diagnostic accuracy meta-analysis of the Whooley questions recommended by the 2010 NICE guidelines (25) to be used when depression is suspected. This recommendation, however, was based on limited evidence. The Canadian Network for Mood and Anxiety Treatments (CANMAT) Clinical Guidelines for the Management of Adults with Major Depressive Disorder published in 2016 and the US Preventive Services Task Force for Depression (USPSTF) now recommend that the Whooley questions are used to screen in primary and secondary care settings 'in individuals with risk factors when there are available resources and services for subsequent diagnostic assessment and management.' (10,136). These recommendations have been informed by the accuracy of the Whooley questions reported by Paper 4. (10) This paper provides strong evidence that the Whooley questions are consistently good at ruling out depression.

Paper 4 has been cited in 53 peer reviewed journal articles according to Web of Science as of 7 May 2024 and has 105 Google Scholar citations. The majority of these papers used our study to make decisions about using the Whooley questions in their research, based on the quality of, or the diagnostic properties reported by our review. (129,137–150)

The developer of the instrument herself (Professor Mary Whooley) cited this paper (rather than her own validation study) and paper 3 when she reported the summary properties of the two ultra-brief screening tools for depression (the Whooley questions and the PHQ-2, respectively). Her paper, “Screening for depression - A tale of two questions” , includes an acknowledgment of the quality of our work. (151) Furthermore, this paper was also positively reviewed (alongside our other meta-analyses published prior to 2016, when these methodological reviews were carried out) in three methodological reviews of the quality of meta-analyses of diagnostic accuracy of depression screening tools. (11,12,152)

9.2 Implications of research findings for policy and practice

This review of the diagnostic accuracy of the Whooley questions provides strong evidence of consistent high sensitivity and moderate specificity for the two questions, across a range of settings among different populations. Although the modest specificity means that some people who score positively will not meet diagnostic criteria for depression, the test retains value in its ability to eliminate MDD. The introduction of a help question appeared to improve specificity when used as second tier test in one study, though evidence of its performance remains sparse and contradictory.

As highlighted above, based on the findings of this review, the Canadian and US guidelines identified the Whooley questions as ‘an effective and simple approach for screening in clinical practice’. They advise that an answer of “yes” to either question requires a more detailed assessment. (10,136)

9.3 Implications for future research

This paper makes important recommendations on improving the quality and reporting of validation studies of ultra-brief SCFI. As identified in previous papers, the QUADAS-2 ratings indicate that there are a number of limitations of the primary studies and often details about key methodological criteria were not reported.

The paper suggests a number of research recommendations. Future diagnostic validation studies should report sufficient detail on the method to permit an assessment of key methodological criteria. Subsequent reviews of the Whooley would benefit from a more consistent method of referring to the Whooley in primary studies. The review recommended the use of the term 'Whooley questions' and avoidance of the term 'PHQ-2'. Although the PHQ-2 shares similarities with the Whooley questions, the PHQ-2 asks about a different time frame and uses a different scoring system. We recommended that future studies should refer to Whooley in the title or abstract in order to facilitate future reviews of the measure. Although it may seem trivial, this paper emphasises the importance of using the exact name of the instrument when a validation study is conducted, in order to enable future diagnostic accuracy reviews.

10 Paper 5

Pocklington, C, McMillan, D, Gilbody, SM & Manea, LE 2016, 'The diagnostic accuracy of brief versions of the Geriatric Depression Scale: a systematic review and meta-analysis', *International Journal of Geriatric Psychiatry*, Volume 31, Issue 8 pp. 837 - 857

Paper 5 aimed to establish the diagnostic accuracy of brief versions of the widely used Geriatric Depression Scale (GDS). These are presented in the Results section of the paper.

10.1 How does this study contribute to the body of knowledge?

This paper provides updated information regarding the diagnostic performance of the GDS-15. It is difficult to make firm conclusions because (similarly to the PHQ-9 diagnostic meta-analyses) the pooled results show evidence of selective reporting of cut-off scores; therefore, these findings should be interpreted cautiously. Selective reporting of cut-off scores leads to the diagnostic accuracy of the screening instruments being exaggerated because results for cut-off scores that perform less well are not reported.

Briefer versions of the GDS may have more clinical appeal owing to the time restraints faced in clinical practice. Unfortunately, meta-analyses were not possible for briefer versions because of an inadequate number of primary studies. Several briefer versions of the GDS were found; GDS-1, GDS-4, GDS-5, GDS-7, GDS-8 and GDS-10. However there was inconsistency in the items that contributed to these briefer versions and no standardised cut-off scores; therefore, based on the evidence summarized in this review, none of these versions can be currently recommended for wider use.

Paper 5 has been cited in 162 peer reviewed journal articles according to Web of Science as of 7 May 2024 and has 239 Google Scholar citations. The majority of these papers used

our study to make decisions about using the GDS in their research, based on the quality of, or the diagnostic properties reported by our review. (153–170)

Similarly to the previous papers that form this thesis, this paper was cited as one of the systematic reviews and DTA meta-analyses that fulfilled the criteria on the highest number of AMSTAR items. (12)

10.2 Implications of research findings to policy and practice

Depression in older adults is often under-recognised despite it being the most common mental health illness in this age group. An increasing older adult population highlights the need for improved diagnostic rates. Brief versions (15 items or less) of the GDS, which are suitable for busy clinical practice, could improve detection rates. This paper showed that GDS-15 has a sensitivity of 0.77 and a specificity of 0.89 at the recommended cut-off score of 5. The sensitivity reported by this review is lower than that reported by previous reviews whilst the specificity is higher. At a cut-off score of 4 diagnostic data was more favorable; sensitivity was 0.88 and specificity was 0.86, which resulted in a greater DOR compared to a cut-off score of 5 (42.05 versus 27.28).

A significant issue raised by this study was that not all studies measured cognitive functioning, and some excluded patients with cognitive impairment. Obviously, the presence of cognitive impairment may substantially affect the diagnostic accuracy of a depression measure in an older adult population. We recommended that future studies of the GDS may want to report its diagnostic performance separately for samples including and then excluding people with cognitive deficits. The protocol of an IPD meta-analysis of different versions of GDS has specifically followed this recommendation. (171)

Meta-analyses could not be performed for any briefer versions of the GDS because of an inadequate number of studies for the different cut-off scores reported, and due to the

variability of items that made up the different versions. Therefore, briefer versions should not be used until these issues can be explored further.

10.3 Implications for future research

This paper makes important recommendations on improving the quality and reporting of validation studies of ultra-brief SCFIs. As identified by previous papers, the QUADAS-2 ratings indicate a number of methodological issues. The QUADAS-2 domain of 'index test' (please see the rated domains in table 2 in the paper) was particularly concerning and the overall risk of bias of the 'index test' did influence diagnostic performance, as discussed in the paper. Bias concerning the QUADAS-2 domain of 'patient selection' did not influence pooled diagnostic data. For the domain of 'flow/timing' an interval of more than two weeks between administration of the GDS and reference test did influence pooled diagnostic data; specificity increased and sensitivity fell when meta-analysis was re-run excluding primary studies where risk was rated as 'high' or 'unclear'.

As previously highlighted, primary studies on the diagnostic accuracy of the different versions of the GDS have been limited by (1) small samples; (2) the selective reporting of results for cut-offs when they perform well in a given sample, but not when they perform poorly; (3) the inclusion of patients already known by clinicians to have depression and (4) the inability to conduct subgroup analyses (e.g. different age groups, dementia diagnosis, care settings) due to small sample sizes. These concerns, explicitly raised in this paper have informed another IPDMA protocol for a study currently in progress, that aims to establish the DTA of various versions of the GDS using IPD (171)

Knowledge concerning the diagnostic accuracy of versions of the GDS with fewer than 15 items is currently limited. A range of items was used to contribute to these brief versions; therefore, we recommend that an unhelpful proliferation of different versions with limited accuracy data should be avoided.

11 Paper 6

Manea, LE, Boehnke, JR, Gilbody, SM, Moriarty, AS, & McMillan, D 2017, 'Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis' *BMJ open* 7, pp. 1 - 23

Paper 6 examined a methodological issue identified in the PHQ-9 diagnostic accuracy meta-analyses. Based on a post-hoc observation, we investigated whether an allegiance effect is found that leads to an increased diagnostic performance in diagnostic validation studies that were conducted by teams connected to the original developers of the PHQ-9 (the 'non-independent' studies). Please refer to the results section of the paper for findings.

11.1 How does this study contribute to the body of knowledge?

This is the first systematic examination of a possible 'allegiance' or authorship effect in the validation of SCFI for a common mental disorder. The analyses showed that diagnostic studies conducted by independent researchers had lower sensitivity paired with similar specificity compared to studies that were classified as non-independent. This conclusion held for both the algorithm and cut-off 10 studies.

Previous research has proposed several possible explanations for the allegiance effect (172–174). One possibility is the bias that may serve to inflate the performance of a test when evaluated by those who have developed it. However, before concluding that the differences are due to this effect, it is important to explore and rule out alternative explanations.

This paper explored a range of possible alternative explanations for the observed 'allegiance' effect including both differences in study characteristics and study quality. A number of potential differences were found that offered potential alternative

explanations, unconnected to allegiance effects. These included the greater use of the SCID in the studies rated as non-independent in both the algorithm and the cut-off 10 studies. Interestingly an IPD meta-analysis subsequently carried out (104) identified that, based on results from semi-structured interviews only (including SCID), the PHQ-9 was more sensitive, so could explain some of this difference. The non-independent algorithm studies were also more likely to use an appropriate translation of the PHQ-9 and were also more likely to ensure that the index and reference test were conducted within two weeks of each other, both of which may be associated with an improvement in observed diagnostic performance. The majority of studies in both meta-analyses did not provide clear statements about potential conflict of interest and/or funding; however, the newer studies were more likely to provide such statements, which may reflect increasing transparency in this area of research.

This paper has the latest publication date (September 2017) in the group of papers included in this thesis and has been cited 25 times. It is clear from these citations that our work has started an important conversation about the needed scrutiny on potential allegiance bias in diagnostic test accuracy research – see editorial on ‘Spin, Bias, and Clinical Utility in Systematic Reviews of Diagnostic Studies’ (175), but also other citing papers (176,177).

11.2 Implications for future research

Although it has been suggested that allegiance effects may play a role in the validation of psychological SCFI (178) systematic evaluations of this hypothesis are rare and studies that acknowledge potential allegiance effects in such studies mainly come from forensic psychology and psychiatry backgrounds. (178–181) Diagnostic validation studies are designed to establish the sensitivity and specificity of a SCFI, which are used in practice to differentiate cases from non-cases or to decide about whether further assessment or treatment is indicated or will be offered. An allegiance effect in such studies would be

seen in systematically higher sensitivities or specificities if the original authors were part of the study team. Such a bias would have a detrimental effect on practice through promising over-optimistic accuracy of the SCFI or in evaluating the cost-effectiveness of the measure in a screening or case-finding context.

Conflicts of interest are an important area of investigation in medical and behavioural research, particularly due to concerns about results being influenced by industry sponsorship. Future diagnostic validity in this area should present clear statements about potential conflicts of interest and funding, particularly relating to the development of the instrument under evaluation.

12 Final remarks and recommendations

This integrative chapter has summarised the aims, objectives, methodology, results and conclusions from contributing papers. It has clarified how they form a coherent body of work and make a significant and original contribution to knowledge and understanding. The integrative chapter also specified the candidate's contribution to each of the papers.

The portfolio of work presented here aimed to address a number of evidence gaps regarding the performance of the most popular brief and ultra-brief SCFIs for MDD in various non-psychiatric settings. The instruments evaluated in this thesis are widely used in various clinical settings (for instance, in the NHS Improving Access to Psychological Therapies (IAPT) service, the cut-off points recommended by the developers of the PHQ-9 are used to define caseness for service use).(182)

This work has highlighted that, apart from the Whooley questions, the initial validation studies of these SCFIs have reported better performance. Clinicians and policy makers should therefore be aware that these instruments might be less accurate for screening or case-finding purposes than originally reported. Moreover, the recommended cut-off points were not necessarily the best performing, and most papers included in this thesis highlighted that a universal cut-off point may not be appropriate in every setting. Overall, this portfolio of work has highlighted ongoing uncertainty regarding the performance of these instruments in various settings, largely maintained by the generally poor quality of primary validation studies and widespread practices like selective reporting of cut-off points. The extent of these practices has first been highlighted by the work included in this thesis and has set the scene for further in-depth analysis of the impact on diagnostic performance of diagnostic tests (for instance the IPD meta-analyses carried out by the research group led by Professor Thombs).

There are a number of policy, practice and research recommendations that can be drawn

from this work. These are summarised below.

12.1 Policy and practice recommendations

This body of work aimed to review the diagnostic accuracy of the above instruments. The results presented in the included papers could add to the body of evidence to inform the next set of NICE guidelines for identifying and managing depression. The current NICE guidelines, based on the limited evidence available when the guidelines were issued, identified that the Whooley questions, PHQ-9 and the GDS may perform better as SCFIs for depression. The guidelines also reported that, when compared to the Whooley questions, the PHQ-9 and GDS-15 had better specificity but not as much sensitivity. (25) The work included in this thesis provides more in-depth understanding of the performance of these instruments. The following recommendations, summarized in box 1, are based on significant findings of this body of work, that further map our knowledge and uncertainties of the performance of the best performing SCFIs for depression, as identified by the current NICE guidelines.

Box 1 Main clinical and practice recommendations

- The accuracy of the PHQ-9 diagnostic algorithm does not compare favourably to that of the PHQ-9 using the standard cut-off of ≥ 10 , therefore it is less useful to screen for MDD.
- The cut point of ≥ 10 for the PHQ-9 represents a better diagnostic performance for screening purposes or where a high sensitivity is needed, and is the cut-off score that maximizes sensitivity and specificity.
- The PHQ-2 at ≥ 2 , rather than at the recommended cut-off point of ≥ 3 , has better value in ruling out depression and 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.
- The Whooley questions are clearly good at ruling out depression and appear to be the best primary test for patients raising concerns.
- GDS-15 has a sensitivity of 0.77 and a specificity of 0.89 at the recommended cut-off score of 5. Briefer versions of the GDS may have more clinical appeal owing to the time restraints faced in clinical practice, but there is not enough evidence for their use.

High levels of heterogeneity were found for almost all instruments, which remains an important limitation of the currently available evidence. One potential explanation of heterogeneity is that the tests perform differently in different populations. Subgroup analyses evidenced that the diagnostic properties of the tests differed in various populations.

These reviews also highlight an important finding – that the same single threshold might not be appropriate in all settings. The work summarised in this thesis aimed to explore whether different cut-off points perform differently in different settings. However due to selective reporting of cut-off points no firm conclusions could be drawn. Most studies selectively reported some cut-off points but not others without stating the reason for this.

In summary, the body of work presented here has raised important questions about the accuracy of recommended cut-off points, whether a single cut-off point is appropriate in all settings, as well as important methodological issues related to validation studies of SCFIs. Therefore, it is a very useful resource for clinicians or researchers using these instruments and informs their choice of scoring depending on the population screened or setting. It also identified and explored in more depth important methodological issues (discussed below) that constitute useful information for researchers and research commissioners interested in developing and evaluating SCFIs.

As mentioned above, the findings presented in these reviews have already informed national guidelines in Canada and the US, and will likely be reviewed in order to inform the body of evidence for the next set of NICE guidelines for identifying and managing depression. Currently, national guidelines from Canada and the UK advise against routine screening for depression due to the lack of evidence of benefit from well conducted randomized controlled trials, concerns about high false positive rates, overdiagnosis, and significant resource use and healthcare costs (33,34) It is unclear at this stage whether using any of the SCFIs recommended by current NICE guidelines and examined in this

thesis, even the better performing cut-off points or scoring methods (suggested by some of our findings), would maximize the likelihood that screening would successfully improve mental health and minimize unnecessary resource use and adverse outcomes. Robust trials that are sufficiently powered to evaluate the effects of screening across a range of cut-off scores should be carried out in order to answer this question. (104)

12.2 Research recommendations

The reviews included in this thesis consistently highlight the need to improve the quality of research in this area (DTA), as well as the quality of reporting of primary validation studies, and make important recommendations on how to improve the identified shortcomings in current research practice.

The QUADAS-2 assessment identified variability in the quality of primary validation studies, and in all systematic reviews included, only a small number of studies were rated as at low risk of bias across all domains. The lack of detail in the reporting of studies made it consistently difficult to assess some of the QUADAS-2 criteria.

Only by performing systematic rigorous reviews of several widely validated instruments systemic issues in research practice have become apparent. Researchers using different or novel methodologies (like IPD) further investigated various issues raised by the work included in this thesis. For instance, a significant issue identified is the design used by some papers in which participants who were more likely to be depressed were also more likely to receive the reference standard, which may have introduced a partial verification bias. Also, reporting the percentage of patients with scores above cut-off thresholds in screening questionnaires for depression as disorder prevalence can substantially overestimate prevalence and generate inaccurate epidemiological data. (123)

Some studies have selectively reported cut-off points and this issue was first highlighted by the findings of a paper we published in 2012 and revisited in Paper 2. It is possible that there is a relationship between the observed performance of the examined instruments at a particular cut-off point and the likelihood that it is reported. Future studies should report the performance of the SCFIs 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 2x2 tables that permit the calculation of the full range of accuracy statistics. Future studies should report sufficient information to ensure that a 2x2 table can be reconstructed.

As described above, the role of screening is to identify previously unknown cases, yet typically most validation studies do not differentiate between previously known and 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 only report the diagnostic performance of the instruments in identifying previously unknown cases.

Box 2 Main research recommendations for future validation studies

- The methodology should be reported in sufficient detail to assess the standard QUADAS-2 criteria.
- Reporting of the performance of all instruments' versions should be available for all cut-off points. Studies should also report sufficient information to ensure that a 2x2 table can be reconstructed from the information reported and this should be made a condition for publication.
- The reported diagnostic performance should reflect the instruments' accuracy in identifying previously unknown cases.
- Only validated translation procedures should be used.
- The name of the validated instrument should be referred to correctly and consistently to (e.g. to ensure distinction between Whooley questions and PHQ-2).
- The diagnostic performance of differently constructed measures should be compared to identify which combination has greatest accuracy (e.g. briefer versions of the GDS).

All pooled estimates reported in these reviews need to 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 and this should be addressed in future research.

An issue not previously explored in diagnostic test accuracy research and methodology has been highlighted in the last paper included in this thesis. An allegiance or authorship effect in the PHQ-9 validation studies could potentially have led to systematically higher sensitivities in validation studies where original authors were part of the team. An important recommendation, that all future meta-analyses of diagnostic validation studies of psychological measures should routinely evaluate the impact of researcher allegiance in the primary studies examined in the meta-analysis, has stemmed from this research.

This work has also highlighted that many studies did not provide clear statements about potential conflict of interest and/or funding and has highlighted the need for future validation should routinely present clear statements about these details.

Despite the growing number of validation studies, the accuracy of diagnostic and epidemiological data is compromised by current research practices. The limitations highlighted by the papers included in this thesis, particularly the generally poor reporting of validation studies, have led researchers to employ different methodologies (e.g. IPD meta-analyses) (13,15,101,183,184), and to further explore the methodological issues we systematically identified in our reviews, significantly advancing knowledge in the area of DTA research.

Systematic reviews and diagnostic meta-analyses offer an opportunity to map our knowledge and uncertainties of the performance of diagnostic tools, and identify many of the biases of the primary studies. In some cases, it may also try to correct some of them but at a minimum, a methodologically robust systematic review can at least highlight the presence of biases.(175) However, suboptimal systematic reviews and meta-analyses can be harmful particularly given the influence these types of studies have acquired. (185)

Therefore, in order to add meaningfully to the existent body of knowledge systematic reviews and DTA meta-analyses have to comply with rigorous methodological standards. The papers included in this thesis were cited as the few identified systematic reviews and DTA meta-analyses that fulfilled the criteria on the highest number of AMSTAR items (11,12,99). Subsequently the reported results have consistently been used in preference to those reported by the original validation studies, by numerous studies. The papers included in this thesis have been cited over 800 times, a clear indicator of the impact this work has made in summarising and understanding the accuracy of the most commonly used depression SCFIs, as well as improving systematic reviews and DTA methodology. The papers included in this thesis have led researchers to further explore our findings (e.g. (13,14,101,122,123,135,171,186), and have significantly advanced this area of research.

13 References

1. Evans DL, Charney DS, Lewis L, Golden RN, Gorman JM, Krishnan KRR, et al. Mood Disorders in the Medically Ill: Scientific Review and Recommendations. *Biol Psychiatry*. 2005 Aug;58(3):175–89.
2. Mitchell AJ, Vaze A, Rao S, Greenberg P, Stiglin L, Finkelstein S, et al. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*. 2009 Aug;374(9690):609–19.
3. (UK) NCC for MH. Depression in Adults with a Chronic Physical Health Problem. Depression in Adults with a Chronic Physical Health Problem: Treatment and Management. British Psychological Society; 2010.
4. MacMillan HL, Patterson CJS, Wathen CN, Feightner JW, Bessette P, Elford RW, et al. Screening for depression in primary care: recommendation statement from the Canadian Task Force on Preventive Health Care. *Can Med Assoc J*. 2005 Jan;172(1):33–5.
5. Palmer SC, Coyne JC. Screening for depression in medical care: pitfalls, alternatives, and revised priorities. *J Psychosom Res*. 2003 Apr;54(4):279–87.
6. Joffres M, Jaramillo A, Dickinson J, Lewin G, Pottie K, Shaw E, et al. Recommendations on screening for depression in adults. *Can Med Assoc J*. 2013 Jun;185(9):775–82.
7. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA*. 1999 Nov;282(18):1737–44.
8. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001 Sep;16(9):606–13.
9. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. *CMAJ*. 2012;184(3).
10. Lam RW, McIntosh D, Wang J, Enns MW, Kolivakis T, Michalak EE, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder. *The Canadian Journal of Psychiatry*. 2016 Sep;61(9):510–23.
11. Rice DB, Shrier I, Kloda LA, Benedetti A, Thombs BD. Methodological quality of meta-analyses of the diagnostic accuracy of depression screening tools. *J Psychosom Res*. 2016 May;84:84–92.
12. Rice DB, Kloda LA, Shrier I, Thombs BD. Reporting completeness and transparency of meta-analyses of depression screening tool accuracy: A comparison of meta-analyses published before and after the PRISMA statement. *J Psychosom Res*. 2016;87:57–69.
13. Levis B, Benedetti A, Levis AW, Ioannidis JPA, Shrier I, Cuijpers P, et al. Selective Cutoff Reporting in Studies of Diagnostic Test Accuracy: A Comparison of Conventional and Individual-Patient-Data Meta-Analyses of the Patient Health Questionnaire-9 Depression Screening Tool. *Am J*

- Epidemiol [Internet]. 2017 [cited 2020 Apr 15];185(10):954–64. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28419203>
14. Levis B, Benedetti A, Riehm KE, Saadat N, Levis AW, Azar M, et al. Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews. Vol. 212, *British Journal of Psychiatry*. Cambridge University Press; 2018. p. 377–85.
 15. Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. *The BMJ*. 2019 Apr 9;365.
 16. Ustün TB, Ayuso-Mateos JL, Chatterji S, Mathers C, Murray CJL. Global burden of depressive disorders in the year 2000. *Br J Psychiatry*. 2004 May;184:386–92.
 17. Gensichen J, von Korff M, Peitz M, Muth C, Beyer M, Güthlin C, et al. Case management for depression by health care assistants in small primary care practices: a cluster randomized trial. *Ann Intern Med*. 2009 Sep;151(6):369–78.
 18. Ustün TB, Ayuso-Mateos JL, Chatterji S, Mathers C, Murray CJL. Global burden of depressive disorders in the year 2000. *Br J Psychiatry*. 2004 May;184:386–92.
 19. Spijker J, Graaf R, Bijl R V., Beekman ATF, Ormel J, Nolen WA. Functional disability and depression in the general population. Results from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Acta Psychiatr Scand*. 2004 Sep;110(3):208–14.
 20. McCrone P, Dhanasiri S, Patel A, Knapp M, Lawton-Smith S. Paying the Price: the cost of mental health care in England in 2026 - McCrone, Dhanasiri, Patel, Knapp, Lawton-Smith - The King's Fund, May 2008. 2026;
 21. Gerber PD, Barrett JE, Barrett JA, Oxman TE, Manheimer E, Smith R, et al. The relationship of presenting physical complaints to depressive symptoms in primary care patients. *J Gen Intern Med*. 7(2):170–3.
 22. Cassano P, Fava M. Depression and public health: an overview. *J Psychosom Res*. 2002 Oct;53(4):849–57.
 23. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001 Sep;16(9):606–13.
 24. (UK) NCC for MH. Depression. Depression: The Treatment and Management of Depression in Adults (Updated Edition). British Psychological Society; 2010.
 25. NICE. Depression in adults: recognition and Depression in adults: recognition and management management Clinical guideline. Clinical Guideline 90. 2009.
 26. (UK) NCC for MH. Depression in Adults with a Chronic Physical Health Problem. Depression in Adults with a Chronic Physical Health Problem: Treatment and Management. British Psychological Society; 2010.
 27. Siu A. Screening for depression in adults: US Preventive Services Task Force recommendation statement. *JAMA*. 2016;
 28. Berg AO, Allan JD, Frame PS, Homer CJ, Johnson MS, Klein JD, et al. Screening for depression: recommendations and rationale. *Ann Intern Med* [Internet]. 2002 May 21 [cited 2024 May 7];136(10):760–4. Available from: <https://pubmed.ncbi.nlm.nih.gov/12020145/>

29. Siu A. Screening for depression in adults: US Preventive Services Task Force recommendation statement. *JAMA*. 2016;
30. JW W, MA W, CJ M, GE S, CD M, MP P. Screening for Depression: Recommendations and Rationale. *Ann Intern Med*. 2002 May;136(10):760.
31. MacMillan HL, Patterson CJS, Wathen CN, Feightner JW, Bessette P, Elford RW, et al. Screening for depression in primary care: recommendation statement from the Canadian Task Force on Preventive Health Care. *Can Med Assoc J*. 2005 Jan;172(1):33–5.
32. Allaby M. Screening for depression: A report for the UK National Screening Committee (revised Report). 2010.
33. Allaby M. Screening for depression: A report for the UK National Screening Committee (revised Report). 2010.
34. Joffres M, Jaramillo A, Dickinson J, Lewin G, Pottie K, Shaw E, et al. Recommendations on screening for depression in adults. *Can Med Assoc J*. 2013 Jun;185(9):775–82.
35. Gilbody S, Sheldon T, Wessely S. Should we screen for depression? *BMJ*. 2006;332(7548).
36. Brugha TS, Bebbington PE, Jenkins R. A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. *Psychol Med*. 1999 Sep;29(5):1013–20.
37. Spitzer RL, Williams JBW, Gibbon M, First MB. The Structured Clinical Interview for DSM-III-R (SCID): I: History, Rationale, and Description. *Arch Gen Psychiatry*. 1992;49(8):624–9.
38. World Health Organization. Division of Mental Health. Schedules for clinical assessment in neuropsychiatry. Version 2 : manual. World Health Organization; 1994. 331 p.
39. Robins LN, Wing J, Ulrich Wittchen H, Helzer JE, Babor TF, Burke J, et al. The composite international diagnostic interview. An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Annual Review of Addictions Research and Treatment*. 1991;1(C):263–71.
40. Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule: Its History, Characteristics, and Validity. *Arch Gen Psychiatry*. 1981;38(4):381–9.
41. Sheehan D V., Lecrubier Y, Sheehan KH, Janavs J, Weiller E, Keskiner A, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *European Psychiatry*. 1997 Jan 1;12(5):232–41.
42. Lecrubier Y, Sheehan D V., Weiller E, Amorim P, Bonora I, Sheehan KH, et al. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: Reliability and validity according to the CIDI. *European Psychiatry*. 1997 Jan 1;12(5):224–31.
43. Systematic Reviews: CRD"s guidance for undertaking reviews in health care.
44. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. Quadas-2: A revised tool for the quality assessment of diagnostic accuracy

- studies. Vol. 155, *Annals of Internal Medicine*. American College of Physicians; 2011. p. 529–36.
45. Altman DG, Bland JM. Statistics Notes: Diagnostic tests 1: Sensitivity and specificity. *BMJ* [Internet]. 1994 Jun 11 [cited 2021 Mar 3];308(6943):1552. Available from: <https://www.bmj.com/content/308/6943/1552>
 46. (UK) NCC for MH. *METHOD FOR EVIDENCE SYNTHESIS*. 2015;
 47. Altman DG, Bland j. M. Statistics Notes: Diagnostic tests 2: predictive values. *BMJ* [Internet]. 1994 Jul 9 [cited 2024 Apr 20];309(6947):102. Available from: <https://www.bmj.com/content/309/6947/102.1>
 48. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis. Vol. 29, *Intensive Care Medicine*. Intensive Care Med; 2003. p. 1043–51.
 49. Šimundić A-M. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC* [Internet]. 2009 Jan [cited 2024 Apr 20];19(4):203. Available from: </pmc/articles/PMC4975285/>
 50. *Cochrane Handbook for Systematic Reviews of Interventions* | Cochrane Training [Internet]. [cited 2020 Dec 4]. Available from: <https://training.cochrane.org/handbook/current>
 51. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*. 2005 Oct;58(10):982–90.
 52. Knottnerus JA, Van Weel C, Muris JWM. Evaluation of diagnostic procedures. *BMJ*. 2002 Feb 23;324(7335):477–80.
 53. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med*. 2002 May 15;21(9):1237–56.
 54. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. Vol. 59, *Journal of Clinical Epidemiology*. 2006. p. 1331–2.
 55. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol*. 2004 Sep;57(9):925–32.
 56. Harbord R. METANDI: Stata module to perform meta-analysis of diagnostic accuracy. *Statistical Software Components*. 2008 Apr 16;
 57. Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *J Econom*. 2005 Oct 1;128(2):301–23.
 58. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect of publication bias on meta-analyses. *Br Med J*. 2000 Jun 10;320(7249):1574–7.
 59. 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 Sep;58(9):882–93.
 60. Munhoz TN, Nunes BP, Wehrmeister FC, Santos IS, Matijasevich A. A nationwide population-based study of depression in Brazil. *J Affect Disord*. 2016;192:226–33.

61. Liu Y, Wang J. Validity of the Patient Health Questionnaire-9 for DSM-IV major depressive disorder in a sample of Canadian working population. *J Affect Disord.* 2015;187:122–6.
62. Moriarty AS, Gilbody S, McMillan D, Manea L. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): A meta-analysis. *Gen Hosp Psychiatry.* 2015;37(6).
63. Yu Y, Hu M, Liu Z-W, Liu H-M, Yang JP, Zhou L, et al. Recognition of depression, anxiety, and alcohol abuse in a Chinese rural sample: a cross-sectional study. *BMC Psychiatry.* 2016 Apr;16:93.
64. Kawada T. Depression screening by Patient Health Questionnaire in patients with cardiac surgery. *Int J Cardiol.* 2016 Jun;212:355–6.
65. Pedrelli P, Borsari B, Lipson SK, Heinze JE, Eisenberg D. Gender Differences in the Relationships Among Major Depressive Disorder, Heavy Alcohol Use, and Mental Health Treatment Engagement Among College Students. *J Stud Alcohol Drugs.* 2016 Jul;77(4):620–8.
66. Kawada T, MS V. Psychosocial work stressors and depressive symptoms in bank employees. *Occup Med (Chic Ill).* 2016 Jul;66(5):421.1-421.
67. Hardos JE, Whitehead LW, Han I, Ott DK, Kim Waller D. Depression prevalence and exposure to organophosphate esters in aircraft maintenance workers. *Aerosp Med Hum Perform.* 2016 Aug 1;87(8):712–7.
68. Boeckxstaens GE, Drug V, Dumitrascu D, Farmer AD, Hammer J, Hausken T, et al. Phenotyping of subjects for large scale studies on patients with IBS. *Neurogastroenterology & Motility.* 2016 Aug;28(8):1134–47.
69. Belk RA, Pilling M, Rogers KD, Lovell K, Young A. The theoretical and practical determination of clinical cut-offs for the British Sign Language versions of PHQ-9 and GAD-7. *BMC Psychiatry.* 2016 Nov 3;16(1).
70. Deschênes SS, Burns RJ, Pouwer F, Schmitz N. Diabetes complications and depressive symptoms: Prospective results from the montreal diabetes health and well-being study. *Psychosom Med.* 2017;79(5):603–12.
71. Muñoz-Navarro R, Cano-Vindel A, Medrano LA, Schmitz F, Ruiz-Rodríguez P, Abellán-Maeso C, et al. Utility of the PHQ-9 to identify major depressive disorder in adult patients in Spanish primary care centres. *BMC Psychiatry.* 2017 Aug 9;17(1).
72. Gallis JA, Maselko J, O'Donnell K, Song K, Saqib K, Turner EL, et al. Criterion-related validity and reliability of the Urdu version of the patient health questionnaire in a sample of community-based pregnant women in Pakistan. *PeerJ.* 2018;2018(7).
73. Comtesse H, Rosner R. Prolonged grief disorder among asylum seekers in Germany: the influence of losses and residence status. *Eur J Psychotraumatol.* 2019 Jan 1;10(1).
74. Murillo LA, Grekoff GA, Sheffield JC. Assessing the effect of patient to provider language discordance on depression screening utilizing the Patient Health Questionnaire: An epidemiology study. *Fam Pract.* 2018;36(1):27–31.
75. Zhao N, Zhang Z, Wang Y, Wang J, Li B, Zhu T, et al. See your mental state from your walk: Recognizing anxiety and depression through Kinect-recorded gait data. *PLoS One.* 2019 May 1;14(5).

76. Zhang Y, Li N, Zhao Y, Fan D. Painful Diabetic Peripheral Neuropathy Study of Chinese OutPatients (PDN-SCOPE): Protocol for a multicentre cross-sectional registry study of clinical characteristics and treatment in China. *BMJ Open*. 2019 Apr 1;9(4).
77. Pham T, Bui L, Nguyen A, Nguyen B, Tran P, Vu P, et al. The prevalence of depression and associated risk factors among medical students: An untold story in Vietnam. *PLoS One*. 2019;14(8).
78. Lange S, Burr H, Rose U, Conway PM. Workplace bullying and depressive symptoms among employees in Germany: prospective associations regarding severity and the role of the perpetrator. *Int Arch Occup Environ Health*. 2020 May 1;93(4):433–43.
79. Roberts T, Shiode S, Grundy C, Patel V, Shidhaye R, Rathod SD. Distance to health services and treatment-seeking for depressive symptoms in rural India: A repeated cross-sectional study. *Epidemiol Psychiatr Sci*. 2019;
80. Silverberg ND, Iaccarino MA, Panenka WJ, Iverson GL, McCulloch KL, Dams-O'Connor K, et al. Management of Concussion and Mild Traumatic Brain Injury: A Synthesis of Practice Guidelines. Vol. 101, *Archives of Physical Medicine and Rehabilitation*. W.B. Saunders; 2020. p. 382–93.
81. Narziev N, Goh H, Toshnazarov K, Lee SA, Chung KM, Noh Y. STDD: Short-term depression detection with passive sensing. *Sensors (Switzerland)*. 2020 Mar 1;20(5).
82. Yu Y, Liu ZW, Li TX, Zhou W, Xi SJ, Xiao SY, et al. A comparison of psychometric properties of two common measures of caregiving burden: The family burden interview schedule (FBIS-24) and the Zarit caregiver burden interview (ZBI-22). Vol. 18, *Health and Quality of Life Outcomes*. BioMed Central Ltd.; 2020.
83. Angehrn A, Krakauer RL, Carleton RN. The Impact of Intolerance of Uncertainty and Anxiety Sensitivity on Mental Health Among Public Safety Personnel: When the Uncertain is Unavoidable. *Cognit Ther Res*. 2020 Apr 24;
84. Arrieta J, Aguerrebere M, Raviola G, Flores H, Elliott P, Espinosa A, et al. Validity and Utility of the Patient Health Questionnaire (PHQ)-2 and PHQ-9 for Screening and Diagnosis of Depression in Rural Chiapas, Mexico: A Cross-Sectional Study. *J Clin Psychol*. 2017 Sep 1;73(9):1076–90.
85. Yu Y, Liu ZW, Zhou W, Zhao M, Tang BW, Xiao SY. Determining a cutoff score for the family burden interview schedule using three statistical methods. *BMC Med Res Methodol*. 2019 May 8;19(1).
86. Hyland P, Shevlin M, Brewin CR, Cloitre M, Downes AJ, Jumbe S, et al. Validation of post-traumatic stress disorder (PTSD) and complex PTSD using the International Trauma Questionnaire. *Acta Psychiatr Scand*. 2017 Sep 1;136(3):313–22.
87. Saag LA, Tamhane AR, Batey DS, Mugavero MJ, Eaton EF. Mental health service utilization is associated with retention in care among persons living with HIV at a university-affiliated HIV clinic. *AIDS Res Ther*. 2018 Jan 16;15(1).
88. Ludwig VM, Bayley A, Cook DG, Stahl D, Treasure JL, Asthworth M, et al. Association between depressive symptoms and objectively measured daily step count in individuals at high risk of cardiovascular disease in South London, UK: A cross-sectional study. *BMJ Open*. 2018 Apr 1;8(4).

89. Schuster AK, Tesarz J, Rezapour J, Beutel ME, Bertram B, Pfeiffer N. Visual impairment is associated with depressive symptoms-Results from the nationwide German DEGS1 study. *Front Psychiatry*. 2018 Apr 9;9(APR).
90. Yu Y, Liu ZW, Zhou W, Chen XC, Zhang XY, Hu M, et al. Assessment of burden among family caregivers of schizophrenia: Psychometric testing for short-form Zarit Burden Interviews. *Front Psychol*. 2018 Dec 19;9(DEC).
91. Hirschtritt ME, Kline-Simon AH, Kroenke K, Sterling SA. Depression screening rates and symptom severity by alcohol use among primary care adult patients. *Journal of the American Board of Family Medicine*. 2018 Sep 1;31(5):724–32.
92. Yu Y, Zhou W, Liu ZW, Hu M, Tan ZH, Xiao SY. Gender differences in caregiving among a schizophrenia population. *Psychol Res Behav Manag*. 2019;12:7–13.
93. Zhou K, Jia P. Depressive symptoms in patients with wounds: A cross-sectional study. *Wound Repair and Regeneration*. 2016 Nov;24(6):1059–65.
94. Kato T. Relationship between coping flexibility and the risk of depression in Indian adults. *Asian J Psychiatr*. 2016;24:130–4.
95. Kawada T. Depression screening by Patient Health Questionnaire in patients with cardiac surgery. *Int J Cardiol*. 2016 Jun 1;212:355–6.
96. Hardos JE, Whitehead LW, Han I, Ott DK, Waller DK. Depression Prevalence and Exposure to Organophosphate Esters in Aircraft Maintenance Workers. *Aerosp Med Hum Perform*. 2016 Aug;87(8):712–7.
97. Mitchell AJ, Yadegarfar M, Gill J, Stubbs B. 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. *British Journal of Psychiatry Open*. 2016;2(2).
98. Owora AH, Carabin H, Reese J, Garwe T. Summary diagnostic validity of commonly used maternal major depression disorder case finding instruments in the United States: A meta-analysis. *J Affect Disord*. 2016;205:335–43.
99. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol*. 2007 Dec;7(1):10.
100. Nazar BP, Gregor LK, Albano G, Marchica A, Coco G Lo, Cardi V, et al. Early Response to treatment in Eating Disorders: A Systematic Review and a Diagnostic Test Accuracy Meta-Analysis. *European Eating Disorders Review*. 2016;
101. He C, Levis B, Riehm KE, Saadat N, Levis AW, Azar M, et al. The Accuracy of the Patient Health Questionnaire-9 Algorithm for Screening to Detect Major Depression: An Individual Participant Data Meta-Analysis. *Psychother Psychosom*. 2020 Jan 1;89(1):25–37.
102. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. *CMAJ*. 2012;184(3).
103. Kroenke K, Wu J, Yu Z, Bair MJ, Kean J, Stump T, et al. Patient health questionnaire anxiety and depression scale: Initial validation in three clinical trials. *Psychosom Med*. 2016 Jul 1;78(6):716–27.

104. Levis B, Benedetti A, Thombs BD. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. *The BMJ*. 2019 Apr 9;365.
105. Delgadillo J, Kellett S, Ali S, McMillan D, Barkham M, Saxon D, et al. A multi-service practice research network study of large group psychoeducational cognitive behavioural therapy. *Behaviour Research and Therapy*. 2016 Dec 1;87:155–61.
106. Gilbody S, Brabyn S, Lovell K, Kessler D, Devlin T, Smith L, et al. Telephone-supported computerised cognitive-behavioural therapy: REEACT-2 large-scale pragmatic randomised controlled trial. *British Journal of Psychiatry*. 2017 May 1;210(5):362–7.
107. Hinton L, Sciolla AF, Unützer J, Elizarraras E, Kravitz RL, Apesoa-Varano EC. Family-centered depression treatment for older men in primary care: A qualitative study of stakeholder perspectives. Vol. 18, *BMC Family Practice*. BioMed Central Ltd.; 2017.
108. Gandhi RR, Suthar MA, Pal S, Rathod AJ. Anxiety and depression in spouses of males diagnosed with alcohol dependence: a comparative study. *Archives of Psychiatry and Psychotherapy*. 2017;4:51–6.
109. Gibson-Smith D, Bot M, Snijder M, Nicolaou M, Derks EM, Stronks K, et al. The relation between obesity and depressed mood in a multi-ethnic population. The HELIUS study. *Soc Psychiatry Psychiatr Epidemiol*. 2018 Jun 1;53(6):629–38.
110. Pols AD, Adriaanse MC, Van Tulder MW, Heymans MW, Bosmans JE, Van Dijk SE, et al. Two-year effectiveness of a stepped-care depression prevention intervention and predictors of incident depression in primary care patients with diabetes type 2 and/or coronary heart disease and subthreshold depression: Data from the Step-Dep cluster randomised controlled trial. *BMJ Open*. 2018;8(10).
111. Kealy D, Rice SM, Ferlatte O, Ogrodniczuk JS, Oliffe JL. Better doctor-patient relationships are associated with men choosing more active depression treatment. *Journal of the American Board of Family Medicine*. 2019 Jan 1;32(1):13–9.
112. Wang H, Li T, Yuan W, Zhang Z, Wei J, Qiu G, et al. Mental health of patients with adolescent idiopathic scoliosis and their parents in China: A cross-sectional survey. Vol. 19, *BMC Psychiatry*. BioMed Central Ltd.; 2019.
113. Hackett ML, Teixeira-Pinto A, Farnbach S, Glozier N, Skinner T, Askew DA, et al. Getting it Right: validating a culturally specific screening tool for depression (aPHQ-9) in Aboriginal and Torres Strait Islander Australians. *Medical Journal of Australia*. 2019 Jul 1;211(1):24–30.
114. Roberts T, Shrivastava R, Koschorke M, Patel V, Shidhaye R, Rathod SD. “Is there a medicine for these tensions?” Barriers to treatment-seeking for depressive symptoms in rural India: A qualitative study. *Soc Sci Med*. 2020 Feb 1;246.
115. Roberts T, Shidhaye R, Patel V, Rathod SD. Health care use and treatment-seeking for depression symptoms in rural India: An exploratory cross-sectional analysis. *BMC Health Serv Res*. 2020 Apr 6;20(1).

116. Shin C, Kim Y, Park S, Yoon S, Ko YH, Kim YK, et al. Prevalence and associated factors of depression in general population of Korea: Results from the Korea national health and nutrition examination survey, 2014. *J Korean Med Sci*. 2017 Nov 1;32(11):1861–9.
117. Ng S-M, Leng L-L. Major Depression in Chinese Medicine Outpatients with Stagnation Syndrome: Prevalence and the Impairments in Well-Being. *Evid Based Complement Alternat Med*. 2018;2018.
118. Delgadillo J, Saxon D, Barkham M. Associations between therapists' occupational burnout and their patients' depression and anxiety treatment outcomes. *Depress Anxiety*. 2018 Sep 1;35(9):844–50.
119. Unseld M, Vyssoki B, Bauda I, Felsner M, Adamidis F, Watzke H, et al. Correlation of affective temperament and psychiatric symptoms in palliative care cancer patients. *Wien Klin Wochenschr*. 2018 Nov 1;130(21–22):653–8.
120. Marthoenis, Meutia I, Fathiariani L, Sofyan H. Prevalence of depression and anxiety among college students living in a disaster-prone region. *Alexandria Journal of Medicine*. 2018 Dec 1;54(4):337–40.
121. Kong D, Solomon P, Dong XQ. Comorbid Depressive Symptoms and Chronic Medical Conditions Among US Chinese Older Adults. *J Am Geriatr Soc*. 2019;67:S545–50.
122. Thombs BD, Levis B, Rice DB, Wu Y, Benedetti A. Reducing Waste and Increasing the Usability of Psychiatry Research: The Family of EQUATOR Reporting Guidelines and One of Its Newest Members: The PRISMA-DTA Statement. Vol. 63, *Canadian Journal of Psychiatry*. SAGE Publications Inc.; 2018. p. 509–12.
123. Thombs BD, Kwakkenbos L, Levis AW, Benedetti A. Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. *CMAJ*. 2018 Jan 15;190(2):E44–9.
124. O'byrne R, Cherry KM, Collaton J, Lumley MN, Ca R. The Contribution of Positive Self-Schemas to University Students' Distress and Well-being.
125. Tennenhouse LG, Marrie RA, Bernstein CN, Lix LM. Machine-learning models for depression and anxiety in individuals with immune-mediated inflammatory disease , for the CIHR Team in Defining the Burden and Managing the Effects of Psychiatric Comorbidity in Chronic Immunoinflammatory Disease. 2020;
126. Jiang J, Li Y, Mao Z, Wang F, Huo W, Liu R, et al. Abnormal night sleep duration and poor sleep quality are independently and combinedly associated with elevated depressive symptoms in Chinese rural adults: Henan Rural Cohort. 2019;
127. DiGiovanni G, Mousaw K, Lloyd T, Dukelow N, Fitzgerald B, Poh Loh K, et al. Development of a telehealth geriatric assessment model in response to the COVID-19 pandemic. 2020;
128. Bock C, Heitland I, Zimmermann T, Winter L, Kahl KG. Secondary Traumatic Stress, Mental State, and Work Ability in Nurses—Results of a Psychological Risk Assessment at a University Hospital. *Front Psychiatry*. 2020 Apr 27;11.

129. Slavin V, Creedy DK, Gamble J. Comparison of screening accuracy of the Patient Health Questionnaire-2 using two case-identification methods during pregnancy and postpartum.
130. Bartone PT, Homish GG. Influence of hardiness, avoidance coping, and combat exposure on depression in returning war veterans: A moderated-mediation study. 2020;
131. Ishihara M, Harel D, Levis B, Levis AW, Riehm KE, Saadat N, et al. Shortening Self-report Mental Health Symptom Measures through Optimal Test Assembly Methods: Development and Validation of the Patient Health Questionnaire-Depression-4.
132. Van Damme A, Declercq T, Lemey L, Tandt H, Petrovic M. Late-life depression: issues for the general practitioner. *Int J Gen Med.* 2018;11–113.
133. Darwish L, Beroncal E, Sison MV, Swardfager W. Depression in people with type 2 diabetes: Current perspectives. Vol. 11, *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy.* Dove Medical Press Ltd.; 2018. p. 333–43.
134. Basu S, Landon BE, Williams JW, Bitton A, Song Z, Phillips RS. Behavioral Health Integration into Primary Care: a Microsimulation of Financial Implications for Practices. *J Gen Intern Med.* 2017;32(12):1330–71.
135. Levis B, Sun Y, He C, Wu Y, Krishnan A, Bhandari PM, et al. Accuracy of the PHQ-2 Alone and in Combination with the PHQ-9 for Screening to Detect Major Depression: Systematic Review and Meta-analysis. Vol. 323, *JAMA - Journal of the American Medical Association.* American Medical Association; 2020. p. 2290–300.
136. Whooley MA, (USPSTF) UPSTF, AL S, MP P, BN G, JL R, et al. Screening for Depression—A Tale of Two Questions. *JAMA Intern Med.* 2016 Apr;176(4):436.
137. Ginja S, Jackson K, Newham JJ, Henderson EJ, Smart D, Lingam R. Rural-urban differences in the mental health of perinatal women: a UK-based cross-sectional study. *BMC Pregnancy Childbirth.* 2020 Aug 14;20(1).
138. Cowlshaw S, Metcalf O, Stone C, O'Donnell M, Lotzin A, Forbes D, et al. Posttraumatic Stress Disorder in Primary Care: A Study of General Practices in England. *J Clin Psychol Med Settings.* 2020;
139. Furihata R, Saitoh K, Suzuki M, Jike M, Kaneita Y, Ohida T, et al. A composite measure of sleep health is associated with symptoms of depression among Japanese female hospital nurses. *Compr Psychiatry.* 2020 Feb 1;97.
140. Ingram J, Johnson D, Johnson S, O'Mahen HA, Kessler D, Taylor H, et al. Protocol for a feasibility randomised trial of low-intensity interventions for antenatal depression: ADAGIO trial comparing interpersonal counselling with cognitive behavioural therapy. *BMJ Open.* 2019 Aug 1;9(8).
141. Wang EY, Meyer C, Graham GD, Whooley MA. Evaluating Screening Tests for Depression in Post-Stroke Older Adults. *J Geriatr Psychiatry Neurol.* 2018 May 1;31(3):129–35.
142. Petroulia I, Kyriakos CN, Papadakis S, Tzavara C, Filippidis FT, Girvalaki C, et al. Patterns of tobacco use, quit attempts, readiness to quit and self-efficacy among smokers with anxiety or depression: Findings among six countries of

- the EUREST-PLUS ITC Europe Surveys. Vol. 16, Tobacco Induced Diseases. International Society for the Prevention of Tobacco Induced Diseases; 2018.
143. Howard LM, Ryan EG, Trevillion K, Anderson F, Bick D, Bye A, et al. Accuracy of the Whooley questions and the Edinburgh Postnatal Depression Scale in identifying depression and other mental disorders in early pregnancy. *British Journal of Psychiatry*. 2018 Jan 1;212(1):50–6.
 144. Marsay C, Manderson L, Subramaney U. Validation of the Whooley questions for antenatal depression and anxiety among low-income women in urban South Africa. *South African Journal of Psychiatry*. 2017 Apr 11;23(1).
 145. Manea L, Gilbody S, Hewitt C, North A, Plummer F, Richardson R, et al. Identifying depression with the PHQ-2: A diagnostic meta-analysis. *J Affect Disord*. 2016;203:382–95.
 146. Chorwe-Sungani G, Chipps J. Validity and utility of instruments for screening of depression in women attending antenatal clinics in Blantyre district in Malawi. *South African Family Practice*. 2018 Jul 4;60(4):114–20.
 147. Gollan JK, Wisniewski SR, Luther JF, Eng HF, Dills JL, Sit D, et al. Generating an efficient version of the Edinburgh Postnatal Depression Scale in an urban obstetrical population. *J Affect Disord*. 2017;208:615–20.
 148. Almeida OP, Patel H, Kelly R, Ford A, Flicker L, Robinson S, et al. Preventing depression among older people living in rural areas: A randomised controlled trial of behavioural activation in collaborative care. *Int J Geriatr Psychiatry*. 2020;
 149. Ladds E, Redgrave N, Hotton M, Lamyman M. Systematic review: Predicting adverse psychological outcomes after hand trauma. *Journal of Hand Therapy*. 2017 Oct 1;30(4):407–19.
 150. Cowlshaw S, Gale L, Gregory A, McCambridge J, Kessler D. Gambling problems among patients in primary care: A cross-sectional study of general practices. *British Journal of General Practice*. 2017 Apr 1;67(657):e274–9.
 151. Whooley MA. Screening for depression - A tale of two questions. Vol. 176, *JAMA Internal Medicine*. American Medical Association; 2016. p. 436–8.
 152. Rice DB, Kloda LA, Shrier I, Thombs BD. Reporting quality in abstracts of meta-analyses of depression screening tool accuracy: A review of systematic reviews and meta-analyses. Vol. 6, *BMJ Open*. BMJ Publishing Group; 2016.
 153. Carandang RR, Shibanuma A, Kiriya J, Vardeleon KR, Marges MA, Asis E, et al. Leadership and peer counseling program: Evaluation of training and its impact on Filipino senior peer counselors. *Int J Environ Res Public Health*. 2019 Nov 1;16(21).
 154. Baschi R, Luca A, Nicoletti A, Caccamo M, Cicero CE, D'Agate C, et al. Changes in Motor, Cognitive, and Behavioral Symptoms in Parkinson's Disease and Mild Cognitive Impairment During the COVID-19 Lockdown. *Front Psychiatry*. 2020 Dec 14;11:590134.
 155. Madeira T, Peixoto-Plácido C, Sousa-Santos N, Santos O, Alarcão V, Nicola PJ, et al. Geriatric assessment of the Portuguese population aged 65 and over living in the community: The PEN-3S study. *Acta Med Port*. 2020 Aug 1;33(7):475–82.

156. Arakawa Martins B, Barrie H, Visvanathan R, Daniel L, Arakawa Martins L, Ranasinghe D, et al. A Multidisciplinary Exploratory Approach for Investigating the Experience of Older Adults Attending Hospital Services. *Health Environments Research and Design Journal*. 2020;
157. Carandang RR, Shibanuma A, Kiriya J, Vardeleon KR, Asis E, Murayama H, et al. Effectiveness of peer counseling, social engagement, and combination interventions in improving depressive symptoms of community-dwelling Filipino senior citizens. *PLoS One*. 2020;15(4).
158. Lubitz AF, Eid M, Niedeggen M. Psychosocial and Cognitive Performance Correlates of Subjective Cognitive Complaints in Help-Seeking Versus Non-Help-Seeking Community-Dwelling Adults. *J Geriatr Psychiatry Neurol*. 2020 Mar 1;33(2):93–102.
159. Lin L, Jing XC, Lv SJ, Liang JH, Tian L, Li HL, et al. Mobile device use and the cognitive function and depressive symptoms of older adults living in residential care homes. *BMC Geriatr*. 2020 Feb 3;20(1).
160. Makizako H, Tsutsumimoto K, Doi T, Makino K, Nakakubo S, Liu-Ambrose T, et al. Exercise and Horticultural Programs for Older Adults with Depressive Symptoms and Memory Problems: A Randomized Controlled Trial. *J Clin Med*. 2019 Dec 30;9(1):99.
161. Liew TM. Depression, subjective cognitive decline, and the risk of neurocognitive disorders. *Alzheimers Res Ther*. 2019 Aug 9;11(1).
162. Vermunt L, van Paasen AJL, Teunissen CE, Scheltens P, Visser PJ, Tijms BM. Alzheimer disease biomarkers may aid in the prognosis of MCI cases initially reverted to normal. *Neurology*. 2019 Jun 4;92(23):e2699–705.
163. Carandang RR, Shibanuma A, Kiriya J, Asis E, Chavez DC, Meana M, et al. Determinants of depressive symptoms in Filipino senior citizens of the community-based ENGAGE study. *Arch Gerontol Geriatr*. 2019 May 1;82:186–91.
164. Madeira T, Peixoto-Plácido C, Sousa-Santos N, Santos O, Alarcão V, Goulão B, et al. Malnutrition among older adults living in Portuguese nursing homes: The PEN-3S study. *Public Health Nutr*. 2019 Mar 1;22(3):486–97.
165. Taani MH, Siglinsky E, Kovach CR, Buehring B. Psychosocial factors associated with reduced muscle mass, strength, and function in residential care apartment complex residents. *Res Gerontol Nurs*. 2018 Sep 1;11(5):238–48.
166. Kojima Y, Kumagai T, Hidaka T, Kakamu T, Endo S, Mori Y, et al. Characteristics of facial expression recognition ability in patients with Lewy body disease. *Environ Health Prev Med*. 2018 Jul 18;23(1).
167. Dorow M, Stein J, Pabst A, Weyerer S, Werle J, Maier W, et al. Categorical and dimensional perspectives on depression in elderly primary care patients – Results of the AgeMooDe study. *Int J Methods Psychiatr Res*. 2018 Mar 1;27(1).
168. Arakawa Martins B, Barrie H, Dollard J, Mahajan N, Visvanathan R. Older Adults' Perceptions of the Built Environment and Associations with Frailty: A Feasibility and Acceptability Study. *J Frailty Aging*. 2018 Jan 1;7(4):268–71.

169. Balsamo M, Cataldi F, Carlucci L, Padulo C, Fairfield B. Assessment of late-life depression via self-report measures: A review. Vol. 13, *Clinical Interventions in Aging*. Dove Medical Press Ltd.; 2018. p. 2021–44.
170. Moore RC, Straus E, Dev SI, Parish SM, Sueko S, Eyler LT. Development and pilot randomized control trial of a drama program to enhance well-being among older adults. *Arts in Psychotherapy*. 2017 Feb 1;52:1–9.
171. Benedetti A, Wu Y, Levis B, Wilchesky M, Boruff J, Ioannidis JPA, et al. Diagnostic accuracy of the Geriatric Depression Scale-30, Geriatric Depression Scale-15, Geriatric Depression Scale-5 and Geriatric Depression Scale-4 for detecting major depression: Protocol for a systematic review and individual participant data meta-analysis. *BMJ Open*. 2018 Dec 1;8(12).
172. Blair PR, Marcus DK, Boccaccini MT. Is There an Allegiance Effect for Assessment Instruments? Actuarial Risk Assessment as an Exemplar. *Clinical Psychology: Science and Practice*. 2008 Oct;15(4):346–60.
173. Lilienfeld SO, Jones MK. Allegiance Effects in Assessment: Unresolved Questions, Potential Explanations, and Constructive Remedies. *Clinical Psychology: Science and Practice*. 2008 Oct;15(4):361–5.
174. Singh JP, Grann M, Fazel S. Authorship Bias in Violence Risk Assessment? A Systematic Review and Meta-Analysis. Smalheiser NR, editor. *PLoS One*. 2013 Sep;8(9):e72484.
175. Ioannidis JPA. Spin, Bias, and Clinical Utility in Systematic Reviews of Diagnostic Studies. *Clin Chem*. 2020 Jul 1;66(7):863–5.
176. Toner P, Böhnke JR, Andersen P, McCambridge J. Alcohol screening and assessment measures for young people: A systematic review and meta-analysis of validation studies. Vol. 202, *Drug and Alcohol Dependence*. Elsevier Ireland Ltd; 2019. p. 39–49.
177. Simmons J, Wiklund N, Ludvigsson M, Nägga K, Swahnberg K. Validation of REAGERA-S: a new self-administered instrument to identify elder abuse and lifetime experiences of abuse in hospitalized older adults. *J Elder Abuse Negl*. 2020 Mar 14;32(2):173–95.
178. Blair PR, Marcus DK, Boccaccini MT. Is There an Allegiance Effect for Assessment Instruments? Actuarial Risk Assessment as an Exemplar. *Clinical Psychology: Science and Practice*. 2008 Oct;15(4):346–60.
179. Walters GD. The Psychological Inventory of Criminal Thinking Styles and Psychopathy Checklist: Screening version as incrementally valid predictors of recidivism. *Law Hum Behav*. 2009;33(6):497–505.
180. Singh JP, Grann M, Fazel S. Authorship Bias in Violence Risk Assessment? A Systematic Review and Meta-Analysis. Smalheiser NR, editor. *PLoS One*. 2013 Sep;8(9):e72484.
181. Lilienfeld SO, Jones MK. Allegiance Effects in Assessment: Unresolved Questions, Potential Explanations, and Constructive Remedies. *Clinical Psychology: Science and Practice*. 2008 Oct;15(4):361–5.
182. NHS England. NHS England » The Improving Access to Psychological Therapies Manual [Internet]. [cited 2021 Mar 29]. Available from: <https://www.england.nhs.uk/publication/the-improving-access-to-psychological-therapies-manual/>

183. Thombs BD, Benedetti A, Kloda LA, Levis B, Riehm KE, Azar M, et al. Diagnostic accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for detecting major depression in pregnant and postnatal women: protocol for a systematic review and individual patient data meta-analyses. *BMJ Open*. 2015 Oct;5(10):e009742.
184. Thombs BD, Benedetti A, Kloda LA, Levis B, Azar M, Riehm KE, et al. Diagnostic accuracy of the Depression subscale of the Hospital Anxiety and Depression Scale (HADS-D) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *BMJ Open*. 2016 Apr;6(4):e011913.
185. Ioannidis JPA. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. Vol. 94, *Milbank Quarterly*. Blackwell Publishing Inc.; 2016. p. 485–514.
186. Levis B, Yan XW, He C, Sun Y, Benedetti A, Thombs BD. Comparison of depression prevalence estimates in meta-analyses based on screening tools and rating scales versus diagnostic interviews: A meta-research review. *BMC Med*. 2019 Mar 21;17(1).
187. Rohde P, Lewinsohn PM, Klein DN, Seeley JR, Gau JM. Key Characteristics of Major Depressive Disorder Occurring in Childhood, Adolescence, Emerging Adulthood, Adulthood. *Clin Psychol Sci*. 2013 Jan;1(1).
188. Kessing L V. Epidemiology of subtypes of depression. *Acta Psychiatr Scand*. 2007 Feb;115(s433):85–9.
189. (WHO) WHO. The ICD-10 classification of mental and behavioural disorders. ICD-10, the ICD-10 classification of mental and behavioural disorders: 1993.
190. American Psychiatric Association., American Psychiatric Association. DSM-5 Task Force. Diagnostic and statistical manual of mental disorders : DSM-5. 2013. 947 p.

14 Appendix 1: Definitions and classifications of depression

Definitions and classifications of depression

Depression refers to a wide range of mental health problems characterised by the absence of a positive affect (a loss of interest and enjoyment in ordinary things and experiences), low mood and a range of associated emotional, cognitive, physical and behavioural symptoms. Distinguishing the mood changes between clinically significant degrees of depression (for example, major depression) and those occurring 'normally' remains problematic and it is best to consider the symptoms of depression as occurring on a continuum of severity (187). The identification of major depression is based not only on its severity but also on persistence, the presence of other symptoms, and the degree of functional and social impairment. The severity of depression is directly correlated with morbidity and adverse consequences. (187,188)

Mood and affect in a major depressive illness are generally unreactive to circumstances, remaining low throughout the course of each day, although for some people mood can vary, with gradual improvement throughout the day but returns to a low mood on waking.

Depression can present with psychological, behavioural and physical symptoms. Physical and behavioural symptoms include tearfulness, irritability, social withdrawal, an exacerbation of pre-existing pains, pains secondary to increased muscle tension, a lack of libido, fatigue and diminished activity, although agitation is common and marked anxiety frequent. Typically there is reduced sleep and lowered appetite (sometimes leading to significant weight loss), but for some people it is recognised that sleep and appetite are increased. (21) Psychological symptoms include loss of interest and enjoyment in everyday life, and feelings of guilt, worthlessness, lowered self-esteem, loss of confidence, feelings of helplessness and hopelessness, suicidal ideation and attempts at self-harm or suicide. Cognitive changes include poor concentration and

reduced attention, pessimistic and recurrently negative thoughts about oneself and the world, mental slowing and rumination. (22)

Major depression is generally diagnosed when a persistent low mood and an absence of positive affect are accompanied by a range of symptoms, the number and combination make a diagnosis being operationally defined.(189,190)

Depressive disorder as defined by ICD-10

The currently used ICD classificatory system is ICD-11 which came into effect globally from 1 January 2022, after the submission of this thesis therefore the previous revision is used as reference. ICD-10, which was endorsed in May 1990 and came into use in World Health Organisation (WHO) Member States as from 1994, states that a depressive episode is characterised by depressed mood, loss of interest and enjoyment, and reduced energy leading to increased fatigability and diminished activity. (189). The lowered mood tends to vary little from day to day, and is often unresponsive to circumstances, although it may show a characteristic diurnal variation over the course of the day.

Other common symptoms of a depressive episode according to ICD 10 are:

- (a) reduced concentration and attention;
- (b) reduced self-esteem and self-confidence;
- (c) ideas of guilt and unworthiness (even in a mild type of episode);
- (d) bleak and pessimistic views of the future;
- (e) ideas or acts of self-harm or suicide;
- (f) disturbed sleep

(g) diminished appetite.

It is generally accepted that a duration of at least 2 weeks is usually required for diagnosis of depressive episodes of all three grades of severity (mild, moderate or severe), but shorter periods may be reasonable if symptoms are unusually severe and of rapid onset.

Some of the above symptoms may be marked and develop characteristic features that are widely regarded as having special clinical significance. The most common 'somatic' symptoms are: waking in the morning 2 hours or more before the usual time; depression worse in the morning; objective evidence of definite psychomotor retardation or agitation (remarked on or reported by other people); marked loss of appetite; weight loss (often defined as 5% or more of body weight in the past month); marked loss of libido. Usually, this somatic syndrome is regarded as present if about four of these symptoms are definitely present.

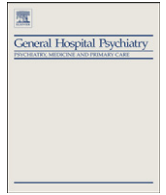
Major depressive episode as defined by DSM-V

The American Psychiatric Association's diagnostic manual defines 'major depressive disorder' (MDD) (also known as recurrent depressive disorder, clinical depression, major depression, unipolar depression, or unipolar disorder) as a mental disorder characterized by an all-encompassing low mood accompanied by low self-esteem, and by loss of interest or pleasure in normally enjoyable activities.

According to formal DSM-V criteria for a Major Depressive Episode the patient should—over a two-week period—experience five or more of the symptoms below, and these must be outside the patient's normal behaviour. Either depressed mood or decreased interest or pleasure must be one of the five (although both are frequently concomitant).(190)

- 15 Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad, empty, hopeless) or observation made by others (e.g., appears tearful). (*Note:* In children and adolescents, can be irritable mood.)
- 16 Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation.)
- 17 Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. (*Note:* In children, consider failure to make expected weight gain.)
- 18 Insomnia or hypersomnia nearly every day.
- 19 Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
- 20 Fatigue or loss of energy nearly every day.
- 21 Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
- 22 Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
- 23 Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.
- 24 The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.
- 25 The episode is not attributable to the physiological effects of a substance or to another medical condition.

Appendix 2: Papers 1-6



A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression



Laura Manea, M.Sc. ^{*}, Simon Gilbody, Ph.D., Dean McMillan, Ph.D.

Hull York Medical School and Department of Health Sciences, University of York, Heslington, York YO105DD, United Kingdom

ARTICLE INFO

Article history:

Received 17 April 2013

Revised 5 September 2014

Accepted 16 September 2014

Keywords:

Depression
Screening
Questionnaire
Psychometrics
Meta-analysis

ABSTRACT

Background: The depression module of the Patient Health Questionnaire-9 (PHQ-9) is a widely used depression screening instrument in nonpsychiatric settings. The PHQ-9 can be scored using different methods, including an algorithm based on *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* criteria and a cut-off based on summed-item scores. The algorithm was the originally proposed scoring method to screen for depression. We summarized the diagnostic test accuracy of the PHQ-9 using the algorithm scoring method across a range of validation studies and compared the diagnostic properties of the PHQ-9 using the algorithm and summed scoring method at the proposed cut-off point of 10.

Methods: We performed a systematic review of diagnostic accuracy studies of the PHQ-9 using the algorithm scoring method to detect major depressive disorder (MDD). We used meta-analytic methods to calculate summary sensitivity, specificity, likelihood ratios and diagnostic odds ratios for diagnosing MDD of the PHQ-9 using algorithm scoring method. In studies that reported both scoring methods (algorithm and summed-item scoring at proposed cut-off point of ≥ 10), we compared the diagnostic properties of the PHQ-9 using these methods.

Results: We found 27 validation studies that validated the algorithm scoring method of the PHQ-9 in various settings. There was substantial heterogeneity across studies, which makes the pooled results difficult to interpret. In general, sensitivity was low whereas specificity was good. Thirteen studies reported the diagnostic properties of the PHQ-9 for both scoring methods. Pooled sensitivity for algorithm scoring method was lower while specificities were good for both scoring methods. Heterogeneity was consistently high; therefore, caution should be used when interpreting these results.

Interpretation: This review shows that, if the algorithm scoring method is used, the PHQ-9 has a low sensitivity for detecting MDD. This could be due to the rating scale categories of the measure, higher specificity or other factors that warrant further research. The summed-item score method at proposed cut-off point of ≥ 10 has better diagnostic performance for screening purposes or where a high sensitivity is needed.

© 2015 Elsevier Inc. All rights reserved.

Depressive disorder is the most common mental health problem in primary health care and medical specialty population [1]. However, recognition of depression in these settings is still low. There is substantial decision uncertainty about the value of screening or case finding for depression in primary care settings. There is, for example, substantial disagreement between different national guidance about the benefits of these strategies. US guidelines recommend a form of screening, offered to all regardless of level of risk if there are appropriate structures and processes in place to manage those identified as depressed [2]. UK NICE guidance, while not recommending this general screening approach, recommends an alternative strategy involving the use of brief case-finding instrument for people deemed at increased risk, such as those with chronic physical health problems [3,4]. In contrast, Canadian guidelines [5] strongly caution against the use of any form of

screening or case finding for depression because of, among other concerns, a lack of evidence about the potential harms of screening. The decision about whether to screen or use case-finding procedures for depression would, according to such guidance, alter as a policy maker crossed a national boundary.

The Patient Health Questionnaire-9 (PHQ-9) is a self-report measure of depression consisting of nine items matching the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (DSM-IV) criteria of major depression. Respondents are asked to rate each of the items on a scale of 0 to 3 on the basis of how much a symptom has bothered them over the last 2 weeks (0=not at all, 1=several days, 2=more than half the days, 3=nearly every day). There are different methods of scoring the PHQ-9 to screen for depression, including an algorithm based on DSM-IV criteria and a cut-off based on summed-item scores. The algorithm method requires a total of at least five symptoms rated as at least 2 (more than half the days), with the exception of the suicidal ideation item, which counts as one of the five symptoms if it is rated as 1 (several days) or above. The algorithm also requires that at least one of

^{*} Corresponding author. Hull York Medical School and Department of Health Sciences, ARRC Building, University of York, Heslington, York YO105DD, United Kingdom.

E-mail address: laura.manea@york.ac.uk (L. Manea).

the symptoms scored as at least 2 is either loss of interest or pleasure or depressed mood. A 10th item was added to the diagnostic part of the PHQ-9 asking patients how difficult the problems identified made it for them to manage work, daily living and relationships [6]. In contrast, the summed-item score simply adds up the scores from each of the items to give a total score ranging from 0 to 27. A cut-off score of 10 or above on the summed-item score has been recommended as a method of screening for major depressive disorder [6].

On a priori grounds, the algorithm scoring method might be expected to be superior to the summed-item method because the algorithm matches the DSM-IV criteria for diagnosing major depression that are contained in the gold standard against which the performance of PHQ-9 is to be assessed (i.e., requirement that a core symptom is present, symptoms with the exception of suicidal ideation occur at a specified frequency). In contrast, the cut-off score does not map directly onto diagnostic criteria. The early validation studies of the PHQ-9, however, indicated that the summed-item method may, in fact, be more suitable than the algorithm as a screening or case-finding tool, primarily because of the low sensitivity of the algorithm method. Data from the PHQ-9 primary care study showed that the algorithm had a sensitivity of 73% and a specificity of 98% [7]. In the validation study of the summed-item method, a score of ≥ 10 had a sensitivity of 88% and a specificity of 88% for major depressive disorder (MDD) [6].

Perhaps for this reason, the summed-item scoring method has come to dominate the way in which the PHQ-9 is used to screen for depression, with the algorithm falling into disuse. However, the rejection of the algorithm scoring strategy may be premature on the basis of the early validation studies alone and in the absence of a comprehensive analysis of all of the relevant studies to date in this area, particularly given that, a priori, the algorithm method may be expected to be superior. The aim of the current diagnostic meta-analysis is to examine the diagnostic properties of the PHQ-9 using the algorithm scoring method and to compare it directly with the summed-item scoring method.

1. Methods

In this study, we included all studies of the PHQ-9 that used the algorithm scoring method to screen for MDD, in any setting and any population. We used systematic review and meta-analytic techniques to summarize the diagnostic properties of the PHQ-9 for MDD using the algorithm [8,9]. Where studies reported both the accuracy of the algorithm scoring method and the summed-item scoring method at the standard cut-off point of ≥ 10 , we extracted data on both so that their diagnostic performance could be compared. The systematic review methods used in this review have followed the guidelines and recommendations stipulated by the Centre for Reviews and Dissemination [10]. We performed a diagnostic systematic review of the available literature using bivariate meta-analysis methods [11–13].

1.1. Literature search

In order to capture relevant studies reporting the ability of the PHQ-9 to detect MDD, we searched the databases EMBASE, MEDLINE and PsycINFO from 1999 (when Patient Health Questionnaire was first developed) to August 2013 using the terms PHQ or patient health questionnaire. We aimed to develop a maximally sensitive search to identify all studies that had used the PHQ-9. This search (using the terms PHQ/PHQ\$/PHQ-9 or Patient Health Questionnaire) would identify references to the PHQ-9 in the title or abstract. We used the same search strategy that we used in a previous systematic review that identified validation studies of the PHQ-9 for various cut-off points [14]. The full search strategy is presented in Appendix 1.

For each study that met full inclusion criteria, we manually searched the reference lists and performed a reverse citation search in Web of Science to identify additional studies. We corresponded with the authors of original studies to obtain unpublished data where needed. We also

contacted the authors of unpublished studies and conference abstracts in an attempt to minimize publication bias. We applied no publication status or language restrictions.

1.2. Inclusion–exclusion criteria

The following inclusion–exclusion criteria were used:

Population: Any population or setting was included. *Instrument:* We included studies that used the PHQ-9 scored using the algorithm. *Comparison (reference standard):* The accuracy of the PHQ-9 had to be assessed against a recognized gold-standard instrument for the diagnosis of either *Diagnostic and Statistical Manual (DSM)* or *International Classification of Diseases* criterion for major depression. Studies were included if the diagnoses were made using a standardized diagnostic structured interview schedule [e.g., Mini International Neuropsychiatric Interview (MINI), Structured Clinical Interview for DSM Disorders (SCID)]. Unguided clinician diagnoses with no reference to a standard structured diagnostic schedule or comparisons of PHQ-9 with other self-report measures 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 algorithm. *Study design:* Any design. *Additional criterion:* We avoided double counting of evidence by ensuring that only one study of those which reported overlapping datasets in different journals was included in the meta-analysis. 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.

From the electronic searches, the full-text articles for the studies that met these inclusion criteria were retrieved. The final selection was made after examining the full texts. Fig. 1 presents the number of the studies found at each step.

1.3. Data abstraction

We used a standardized data collection form to collect information on the studies. The study features that we extracted and coded sample characteristics (country, setting, age, gender), sample size and percentage with major depression according to the gold standard; information on the PHQ-9 (method of administration, language); and details of the reference standard. Where necessary, authors were contacted to provide clarification. We recorded accuracy data in contingency tables for the algorithm scoring method and, if reported, the cut-off point of 10 using the summed-item scoring method.

1.4. Quality assessment

Quality assessment was conducted at the study level and used criteria based on the QUADAS-II [15]. The QUADAS-II 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 “Was there an appropriate interval between the index test and reference standard?”, we defined an appropriate interval as less than 2 weeks in keeping with how this item has been applied in previous diagnostic test accuracy studies of depression [16].

We added four additional questions that were applied to studies using translated versions of the PHQ-9 and reference test. For translations of the PHQ-9, we asked whether appropriate translation methods were used and whether psychometric properties of the translated version were reported. The same two questions (appropriate translation,

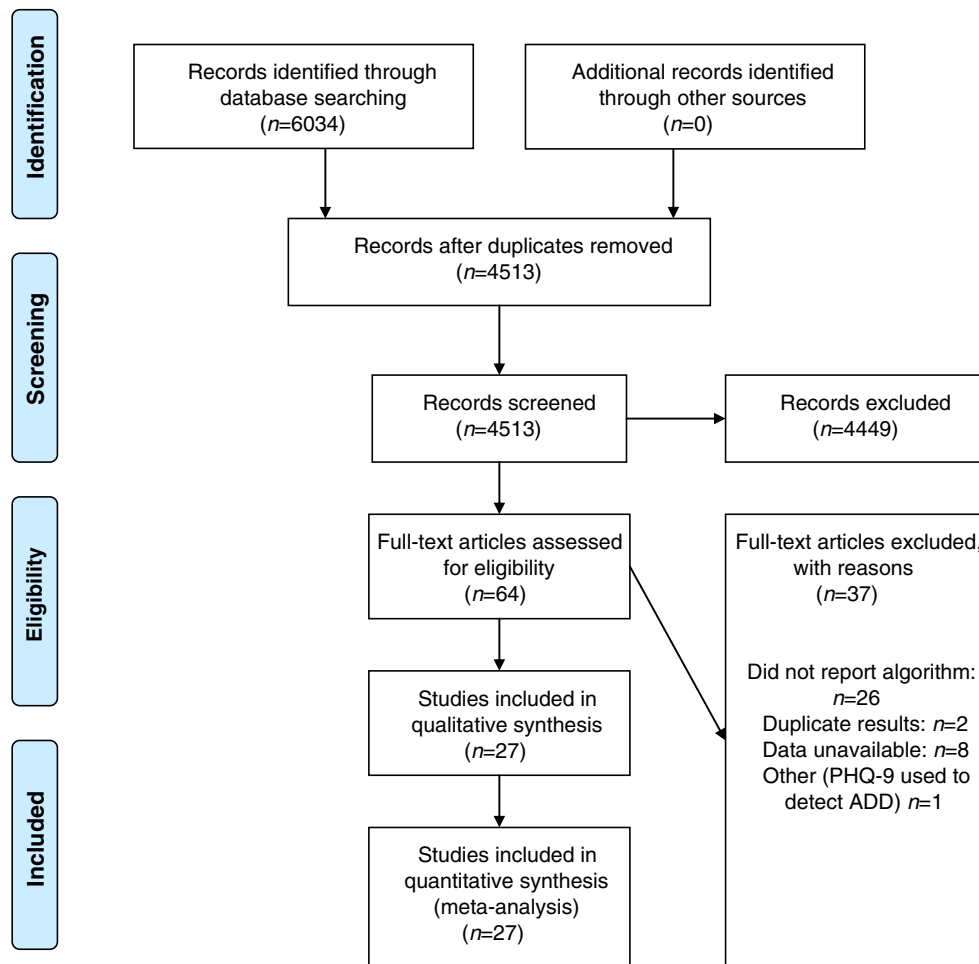


Fig. 1. PRISMA flowchart – search and selection of included diagnostic accuracy studies for systematic review.

psychometric properties) were also applied to any translated version of the reference test.

1.5. Data synthesis and statistical analysis

We constructed 2×2 tables and constructed contingency tables with true positive, true negative, false positive and false negative results.

We performed a bivariate diagnostic meta-analysis to obtain pooled estimates of specificity, sensitivity, likelihood ratios, diagnostic odds ratios (DORs) and their associated 95% confidence intervals (CIs). The bivariate model is a 2-level model that takes into account the precision by which differences in sensitivity and specificity have been calculated while incorporating and estimating the amount of between-study variability in both sensitivity and specificity [17].

1.6. Heterogeneity

It is essential to evaluate heterogeneity (clinical and methodological differences between the studies) in a meta-analysis. Statistical heterogeneity may be caused by known clinical differences between studies or by methodological differences, or it may be related to unknown or unrecorded study characteristics [18].

We measured the between-study heterogeneity using the I^2 statistic of the pooled DOR [19]. I^2 describes the percentage of total variation across studies, which is caused by heterogeneity rather than chance. The I^2 has a greater statistical power to detect clinical heterogeneity when fewer studies are available compared to other measures of heterogeneity. I^2 values of 25% may be considered low; 50%, moderate;

and 75%, high. We explored the causes of heterogeneity where there was significant between-study heterogeneity by visually inspecting the summary receiver operation characteristic curves and identifying the studies that were outside the 95% confidence ellipse. We also undertook a meta-regression analysis of logit DOR using a priori potential sources of heterogeneity entered as covariates in the meta-regression model [12]. We investigated the heterogeneity resulting from sample or study design characteristics by exploring the effects of potential predictive variables [11]. For the sample, we examined the effect of language (translated versus not translated), baseline prevalence of MDD in the screened population, as a proxy measure of the spectrum of severity of disorder within the screened population, and study settings (primary care/community versus general hospital). For study quality, we considered blinding (of the assessor to the results of the PHQ-9 as well as the gold standard) and whether the studies avoided a case-control design or an artificially inflated base rate of MDD. If these items were important sources of heterogeneity, then they would be predictive in a meta-regression analysis and would reduce the level of between-study heterogeneity in the meta-regression model.

Analyses were conducted using STATA version 12, with the metandi, metabias and metareg user-written commands.

2. Results

The initial search identified 4513 unique citations (6034 citations before de-duplication). Of these citations, 64 met initial inclusion criteria and were selected for further screening of the full article. Of the 64 citations, 27 met final stage inclusion criteria [7,20–45].

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. [20]	Country: New Zealand Setting: primary care Age (years): Av.= 49 (range=17–99) Female: 61%	N=2642 Depressed: 6.2%	Administration: not stated Language: English	DSM-IV CIDI
Ayalon et al. [21]	Country: Israel Setting: primary care Age (years): M=75 (S.D.=8.1) Female: 40.5%	N=153 Depressed: 3.9%	Administration: researcher administered Language: Hebrew	DSM-IV SCID
Diez-Quevedo et al. [22]	Country: Spain Setting: medical and surgical tertiary hospitals Age (years): M=43 (S.D.=14.2) Female: 45.6%	N=1003 Depressed: 8.2%	Administration: self-report Language: Spanish	DSM-III-R SCID
Eack et al. [23]	Country: US Setting: community mental health centers for children Age (years): M=39.20 (S.D. 9.63) Female: 100%	N=50 Depressed: 28%	Administration: self-report Language: English	DSM-IV SCID
Fann et al. [24]	Country: US Setting: trauma hospital (inpatients with traumatic brain injury) Age (years): M=42 (S.D.=17.9) Female: 29.1%	N=135 Depressed: 16.3%	Administration: telephone administered Language: English	DSM-IV SCID
Gelaye et al. (2011)	Country: Ethiopia Setting: general hospital Age (years): 34.9 (S.D.=11.6) Female: 63.1%	N=363 Depressed: 12.6%	Administration: researcher administered Language: Amharic	DSM-IV SCAN
Gjerdingen et al. [26]	Country: US Setting: community Age (years): M=29.3 Female: 100%	N=438 Depressed: 4.6%	Administration: telephone or self-report Language: English	DSM-IV SCID
Gräfe et al. (2004)	Country: Germany Setting: psychosomatic walk-in clinics and family practices Age (years): M=41.9 (S.D.=13.8) Female: 67.8%	N=528 Depressed: 29.2% psychosomatic patients; 6.16% medical patients	Administration: self-report Language: German	DSM-IV SCID
Henkel et al. [28]	Country: Germany Setting: primary care Age (years): not reported Female: 74%	N=448 Depressed: 10%	Administration: self-report Language: German	DSM-IV CIDI
Hyphantis et al. [29]	Country: Greece Setting: hospital (rheumatology patients) Age (years): M=54.2 (S.D.=13.5) Female: 74%	N=213 Depressed: 32.4%	Administration: researcher administered Language: Greek	DSM-IV MINI
Inagaki et al. [30]	Country: Japan Setting: general hospital Age whole sample (years): M=73.5 (S.D.=12.3) Female: 59.3%	N=104 out of 511 received MINI Depressed: 7.4%	Administration: researcher administered Language: Japanese	DSM-IV MINI
Khamseh et al. [31]	Country: Iran Setting: diabetes clinic Age (years): M=56.17 (S.D.=9.60) Female: 51.9%	N=185 Depressed: 43.2%	Administration: self-report Language: Persian	DSM-IV SCID
Lamers et al. [32]	Country: The Netherlands Setting: primary care (elderly) Age (years): M=71.4 (S.D.=6.90) Female: 48.2%	N=713 Depressed: 10.7%	Administration: self-report Language: Dutch	DSM-IV MINI
Lotrakul et al. [33]	Country: Thailand Setting: primary care Age (years): M=45.0 (S.D.=14.30) Female: 73.7%	N=279 Depressed: 6.8%	Administration: self-report Language: Thai	DSM-IV MINI
Lowe et al. [34]	Country: Germany Setting: outpatient clinics and family practices Age (years): M=41.7 (S.D.=13.8) Female: 67.1%	N=501 Depressed: 13.2%	Administration: self-report Language: German	DSM-IV SCID
Muramatsu et al. [35]	Country: Japan Setting: primary care and general hospital Age (years): M=43.3 (S.D.=16.4) Female: 59.5%	N=131 Depressed: 28.2%	Administration: self-report Language: Japanese	DSM-IV MINI
Navines et al. [36]	Country: Spain Setting: general hospital (patients with chronic hepatitis C virus) Age (years): M=43.4 (S.D.=10.2) Female: 28.6%	N=500 Depressed: 6.4%	Administration: self-report Language: Spanish	DSM-IV SCID
Persoons et al. [37]	Country: Belgium Setting: hospital (otolaryngology patients) Age (years): M=48.2 (S.D.=12.9) Female: 65.6%	N=268 (97 received MINI) Depressed: 16.5%	Administration: self-report Language: Dutch	DSM-IV MINI

Table 1 (continued)

Study	Sample characteristics (country, setting, age, sex)	Sample size and % depressed	PHQ-2 characteristics	Diagnostic standard
Picardi et al. [38]	Country: Italy Setting: hospital (dermatology inpatients) Age (years): M=37.5 Female: 56%	N= 141 Depressed: 8.5%	Administration: self-report Language: Italian	DSM-IV SCID
Spitzer et al. (1999)	Country: US Setting: primary care Age (years): M=46 (S.D.= 17.2) Female: 66%	N= 3000 (585 received SCID) Depressed: 10%	Administration: self-report Language: English	DSM-III-R SCID
Stafford et al. [39]	Country: Australia Setting: hospital (cardiology patients) Age (years): M=64.1 (S.D.= 10.3) Female: 66%	N= 193 Depressed: 18%	Administration: self-report Language: English	DSM-IV MINI
Thekkumpurath et al. (2010)	Country: UK Setting: hospital (cancer patients) Age (years): M=61 Female: 63%	N= 782 Depressed: 6.3% (of the whole sample)	Administration: not stated Language: English	DSM-IV SCID
Thombs et al. [41]	Country: US Setting: hospital (outpatients with coronary heart disease) Age (years): M=67 (S.D.= 11) Female: 18%	N= 1024 Depressed: 22%	Administration: not stated Language: English	DSM C-DIS
Thompson et al. (2010)	Country: US Setting: patients with Parkinson's disease Age (years): 72.5 (S.D.= 9.6) Female: 42%	N= 214 Depressed: 14%	Administration: self-administered Language: English	DSM-IV SCID
Turner et al. [43]	Country: Australia Setting: stroke patients Age (years): 66.7 (S.D.= 13.1) Female: 47.2%	N= 72 Depressed: 18%	Administration: self-administered Language: English	DSM-IV SCID
van Steenberg-Weijnenburg et al. [44]	Country: The Netherlands Setting: diabetes patients Age (years): M=61.8 (S.D.= 13.6) Female: 48.7%	N= 197 Depressed: 18.8%	Administration: self-administered Language: Dutch	DSM-IV SCID
Zuithoff et al. [45]	Country: The Netherlands Setting: primary care Age (years): M=51 (S.D.= 16.7) Female: 63%	N= 1338 Depressed: 13%	Administration: self-report Language: Dutch	DSM-IV CIDI

Abbreviations: C-DIS, Computerized Diagnostic Interview Schedule; CIDI, Composite International Diagnostic Interview; DSM-III-R, *Diagnostic and Statistical Manual of Mental Disorders, Revised Third Edition*; SCAN, Schedule for Clinical Assessments in Neuropsychiatry.

The remaining 37 were excluded for the following reasons: reference standard diagnosis was not solely major depression ($N=1$), study reported insufficient information to calculate a 2×2 table ($N=8$), studies did not report the diagnostic properties of the PHQ-9 using the algorithm scoring method ($N=26$) and it did not overlap in samples with included studies ($N=2$). The selection of studies is summarized in the PRISMA flowchart [46] in Fig. 1 and further details about the reasons for exclusion are given in Appendix 2.

2.1. Overview of included studies

Table 1 summarizes the characteristics of the included studies. Seven studies were conducted in primary care settings [7,20,21,28,32,33,45]. A further two studies used a combination of a primary care setting and another setting, such as outpatient clinics [34,35]. Sixteen studies recruited from hospital- or outpatient-based medical specialties [22,24,27,29–31,36–43]. Two studies recruited from community samples [23,26].

All of the studies had working age or older adult samples. In the majority of studies, there were more females than males or the samples were entirely female. Mean age ranged from 29.3 years [26] to 75 years [21]. Within these studies, the prevalence of MDD, as diagnosed by the gold-standard tests, ranged between 3.9% [21] and 43.2% [29]. Some of the studies have a high prevalence of major depression because the study design oversampled those who met criteria for major depression or were more likely to meet criteria for major depression (e.g., oversampled those more likely to be depressed on the basis of a high PHQ-9 score).

Eighteen studies stated that a self-report version of the PHQ-9 was used [7,22,23,27,28,31–39,42–45]. In one study, it was administered

over the telephone [24], and in four studies, it was administered by a clinician [21,25,29,30]. In one study, the PHQ-9 was administered either over the phone or was self-reported [27]. The remaining studies did not clearly state the method of administration. Translated versions of the PHQ-9 were used in 16 studies, including Amharic [25], Dutch [32,37,44,45], German versions [27,34], Greek [29], Hebrew [21], Italian [38], Japanese [30,35], Persian [31], Spanish [22,36] and Thai [33].

2.2. Quality assessment

Table 2 summarizes the results of the quality assessment using QUADAS-II. 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 [20,34,45]. The reference standard in Zuithoff et al. [45] assessed major depression over a 6-month timeframe; thus, unlike the PHQ-9, it is not assessing current depression. This may have lowered the observed accuracy of the PHQ-9 in that study. A number of studies had high prevalence rates of major depression because the studies use a design in which participants who are at an increased risk of depression (e.g., those scoring above the threshold on the PHQ-9) were more likely to be given the reference standard.

2.3. Diagnostic properties of the PHQ-9 using diagnostic algorithm

Twenty-seven studies reported the diagnostic properties of the PHQ-9 using the diagnostic algorithm. The pooled sensitivity was 0.58 (CI 0.50–0.66), pooled specificity was 0.94 (CI 0.92–0.96), pooled positive likelihood ratio was 10.81 (CI 7.87–14.86), pooled negative likelihood ratio was 0.43 (CI 0.35–0.52) and DOR was 24.92 (16.73–37.12).

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-9 interpreted blind to reference test	Index test: if translated, appropriate translation	Index test: if translated, psychometric properties reported	Index test: overall risk of bias
Arroll et al. [20]	✓	✓	✓	Low	✓	n/a	n/a	Low
Ayalon et al. [21]	?	✓	✓	Unclear	?	✓	?	Unclear
Diez-Quevedo et al. [22]	x	✓	x	High	?	✓	✓	Unclear
Eack et al. [23]	?	✓	?	Unclear	?	n/a	n/a	Unclear
Fann et al. [24]	x	x	x	High	?	n/a	n/a	Unclear
Gelaye et al. [25]	?	x	?	High	✓	✓	?	Unclear
Gjerdingen et al. [26]	✓	✓	✓	Low	?	n/a	n/a	Unclear
Gräfe et al. (2004)	✓	✓	✓	Low	?	n/a	n/a	Unclear
Henkel et al. [28]	✓	✓	✓	Low	?	n/a	n/a	Unclear
Hyphantis et al. [29]	✓	✓	x	High	✓	?	?	Unclear
Inagaki et al. [30]	✓	x	✓	High	✓	?	?	Unclear
Khamseh et al. [31]	✓	✓	?	Unclear	✓	✓	?	Unclear
Lamers et al. [32]	✓	x	x	High	✓	?	?	Unclear
Lotrakul et al. [33]	x	✓	?	High	✓	✓	?	Unclear
Lowe et al. [34]	x	✓	✓	Low	✓	n/a	n/a	Low
Muramatsu et al. [35]	?	✓	?	Unclear	✓	✓	?	Unclear
Navines et al. [36]	✓	✓	✓	Low	✓	✓	?	Unclear
Persoons et al. [37]	✓	✓	✓	Low	✓	✓	n/a	Unclear
Picardi et al. [38]	✓	✓	✓	Low	✓	?	?	Unclear
Spitzer et al. (1999)	x	✓	✓	High	✓	n/a	n/a	Low
Stafford et al. [39]	✓	✓	✓	Low	✓	n/a	n/a	Low
Thekkumpurath et al. (2010)	x	x	✓	High	✓	n/a	n/a	Low
Thombs et al. [41]	x	✓	?	Unclear	?	n/a	n/a	Unclear
Thompson et al. (2011)	?	✓	✓	Unclear	?	n/a	n/a	Unclear
Turner et al. [43]	?	✓	✓	Unclear	?	n/a	n/a	Unclear
van Steenberg-Weijnenburg et al. [44]	?	✓	✓	Unclear	?	?	?	Unclear
Zuithoff et al. [45]	x	✓	✓	Low	✓	✓	?	Low

Study	Reference test: reference test correctly classifies target condition	Reference test: reference test interpreted blind to PHQ-9	Reference test: if translated, appropriate translation	Reference test: if translated, psychometric properties reported	Reference test: overall risk of bias	Flow/timing: interval of 2 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. [20]	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Ayalon et al. [21]	✓	?	✓	?	Unclear	?	✓	✓	Unclear
Diez-Quevedo et al. [22]	✓	✓	✓	?	Unclear	✓	✓	✓	Low
Eack et al. [23]	✓	?	n/a	n/a	Unclear	?	✓	?	Unclear
Fann et al. [24]	✓	✓	n/a	n/a	Low	✓	✓	x	High
Gelaye et al. [25]	✓	✓	?	?	Unclear	✓	✓	x	High
Gjerdingen et al. [26]	✓	?	n/a	n/a	Unclear	✓	✓	x	High
Fann et al. [24]	✓	✓	n/a	n/a	Low	✓	✓	x	High
Gräfe et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
Henkel et al. [28]	✓	?	n/a	n/a	Unclear	✓	✓	x	High
Hyphantis et al. [29]	✓	✓	?	?	Unclear	✓	✓	x	High
Inagaki et al. [30]	✓	✓	✓	?	Unclear	✓	✓	x	High
Khamseh et al. [31]	✓	✓	✓	?	Unclear	✓	✓	?	Unclear
Lamers et al. [32]	✓	?	?	?	Unclear	?	✓	x	High
Lotrakul et al. [33]	✓	✓	✓	✓	Low	?	✓	x	High
Lowe et al. [34]	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
Muramatsu et al. [35]	✓	✓	✓	✓	Low	✓	✓	?	Unclear
Navines et al. [36]	✓	✓	?	?	Unclear	✓	✓	✓	Low
Persoons et al. [37]	✓	✓	?	?	Unclear	✓	✓	✓	Low
Picardi et al. [38]	✓	✓	✓	?	Unclear	✓	✓	x	High
Spitzer et al. (1999)	✓	✓	n/a	n/a	Low	✓	✓	x	High
Stafford et al. [39]	✓	✓	n/a	n/a	Low	✓	✓	x	High
Thekkumpurath et al. (2010)	✓	✓	n/a	n/a	Low	?	✓	x	High
Thombs et al. [41]	?	✓	n/a	n/a	Unclear	✓	✓	✓	Low
Thompson et al. [42]	✓	?	n/a	n/a	Unclear	✓	✓	x	High
Turner et al. [43]	✓	?	n/a	n/a	Unclear	✓	✓	x	High
van Steenberg-Weijnenburg et al. [44]	✓	x	?	?	High	✓	✓	x	High
Zuithoff et al. [45]	✓	✓	✓	✓	Low	?	✓	✓	Low

✓, criterion met; x, criterion not met; ?, insufficient information to code whether criterion met; n/a, not applicable. If studies reported multiple cut-off points, "threshold pre-specified" is coded as not applicable.

Table 3

Comparative pooled estimates of the PHQ-9 performance using algorithm by setting (primary care versus hospital settings)

Settings	No. of studies	Sensitivity (95% CI)	Specificity (95% CI)	Pooled positive LR (95% CI)	Pooled negative LR (95% CI)	DOR (95% CI)
Primary care	7	0.55 (0.39–0.73)	0.96 (0.94–0.98)	17.69 (10.43–30.00)	0.46 (0.32–0.65)	38.31 (19.27–76.15)
Hospital	17	0.56 (0.46–0.66)	0.93 (0.90–0.95)	9.18 (6.11–13.79)	0.46 (0.37–0.58)	19.78 (11.85–33.00)

Abbreviations: – ve LR, negative likelihood ratio; + ve LR, positive likelihood ratio.

The level of between-study heterogeneity was high (combined DOR $I^2=83.6$). One of the possible reasons for heterogeneity is the various clinical settings in which the PHQ-9 has been validated. On a priori grounds, we conducted subgroup analyses to examine the diagnostic performance of the PHQ-9 in similar clinical settings.

Seven studies were conducted in primary care settings [7,20,21,28,32,33,45] and sixteen studies recruited in hospital- or outpatient-based medical specialties [22,24,27,29–31,36–43]. The DOR using algorithm in hospital settings (DOR=19.78, CI 11.85–33.00) was lower than that in primary care settings (DOR=38.31, 19.27–76.15). Heterogeneity remained high. Studies based on primary care and hospital were again equally heterogeneous (primary care $I^2=82.2\%$; hospital settings $I^2=83.6\%$). For a comparative summary of diagnostic properties of the PHQ-9 in primary care versus hospital settings, see Table 3.

We did not identify a sufficient number of studies (minimum of four studies for a diagnostic meta-analysis) 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 and quality assessment criteria (setting, baseline prevalence of MDD, language, whether the study avoided a case-control design and blinding) were examined as predictors. Out of these variables, only baseline prevalence of MDD was significant ($P=.031$).

2.4. Diagnostic properties of the PHQ-9: comparison of the summed score and algorithm scoring methods

Of the 27 studies, 13 [20,24,26,27,29–31,33,34,39,41,44,45] reported diagnostic properties of the PHQ-9 using both the algorithm and summed-item scoring method at the standard cut-off point of ≥ 10 . Three studies were conducted in primary care [20,33,45]; eight, in hospital settings [24,27,29–31,39,41,44]; one, in community settings [26]; and one, in mixed (psychosomatic walk-in clinics and family practices) settings [27]. Table 4 presents a summary of these results.

In these 13 studies, pooled sensitivity for PHQ-9 using diagnostic algorithm was 0.53 (95% CI 0.42–0.65), pooled specificity was 0.94 (95% CI 0.91–0.96) and DOR was 20.96 (14.10–31.16). When we combined psychometric attributes across studies, we found a moderate level of between-study heterogeneity (combined DOR $I^2=68.7\%$). Pooled sensitivity for PHQ-9 using summed-item scoring methods (cut-off point of 10) was 0.77 (95% CI 0.66–0.85), pooled specificity was 0.85 (95% CI 0.79–0.90) and DOR was 21.53 (15.68–29.58). The level of between-study heterogeneity was $I^2=59.8\%$.

3. Discussion

This systematic review of the diagnostic properties of the PHQ-9 using diagnostic algorithm follows previous recommendations to

summarize diagnostic properties of the PHQ-9 for different scoring methods using a bivariate meta-analysis [47,48]. The review confirmed previous findings that the algorithm method of scoring the PHQ-9 leads to problematically low sensitivity. In both primary care and hospital setting, pooled sensitivity was around 0.55, which is lower than reported in the initial validation study. In either setting, the algorithm method of scoring the PHQ-9 would miss many patients with MDD. However, results should be interpreted with caution because substantial unexplained heterogeneity was found. The only significant variable that was predictive in our meta-regression analysis was the base rate of MDD. In studies directly comparing the algorithm and the standard cut-off point of ≥ 10 of the summed-item scoring method, the summed-item scoring method had a better sensitivity (0.77) and maintained good specificity (0.85); however, caution is again needed in interpreting these results because the level of heterogeneity was substantial.

A possible explanation of the low sensitivity of the algorithm method could lie in the proposed coding strategy, which, with the exception of the suicidal ideation question, determines items scored 2 or 3 as meeting depression criteria, whereas items scored as 1 do not meet criteria. Distinguishing between 1 (several days) and 2 (more than half the days), response categories may be confusing for the respondent. A previous study that explored the psychometric properties of the PHQ-9 concluded that respondents have difficulties differentiating between the two intermediate rating scale categories (several days and more than half the days) and found that the measurement properties of the PHQ-9 can be improved by collapsing rating scale categories [49]. However, there is a substantial body of literature showing that the PHQ-9 score performs very well as a continuous 0- to 27-point scale as well as in ordinal categories (0–4, 5–9, 10–14, 15–19, 20–27). This would be unlikely if there were a substantial number of respondents who equated “several days” with “more than half the days” as representing similar levels of severity. Thus, the degree to which this issue explains the lower specificity of the PHQ-9 algorithm scoring approach should be evaluated in future studies. Also, the findings of Williams et al. should be replicated before collapsing PHQ-9 categories 2 and 3.

The included studies were of variable methodological quality. Some studies 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.

There was some lack of detail in the reporting of studies, which made it difficult to assess some of the QUADAS-II criteria. This was particularly the case for the reporting of whether the reference standard was conducted blind to the PHQ-9. Future studies should make clear statements about the blinding of the reference standard and more

Table 4Pooled estimates of the PHQ-9 performance algorithm versus cut-off point of 10 (studies that reported both scoring methods ($n=6$), 1 study analyzed as 2 separate studies)

Scoring method	No. of studies	Sensitivity (95% CI)	Specificity (95% CI)	Pooled positive LR (95% CI)	Pooled negative LR (95% CI)	DOR (95% CI)
Algorithm	13	0.53 (0.42–0.65)	0.94 (0.91–0.96)	10.20 (7.06–14.72)	0.48 (0.38–0.61)	20.96 (14.10–31.16)
Cut-off 10	13	0.77 (0.66–0.85)	0.85 (0.79–0.90)	5.54 (4.10–7.49)	0.25 (0.17–0.37)	21.53 (15.68–29.58)

Note: *Value could not be estimated.

Abbreviations: – ve LR, negative likelihood ratio; + ve LR, positive likelihood ratio.

generally ensure that the method is reported in sufficient detail to assess the standard QUADAS-II criteria.

There are several limitations to this review. Study selection and data extraction were performed by one author, which may have introduced bias. We did not perform a gray literature search; we cannot, therefore, rule out publication bias. Given that heterogeneity was high, we did not establish funnel plots to examine the potential role of small study and publication bias. We were unable to fully explain the large heterogeneity between studies; consequently, caution should be used when interpreting the results.

The PHQ-9 has emerged worldwide as a popular instrument for depression screening within a variety of settings. Our results show that the algorithm scoring method has a low sensitivity and the cut point of ≥ 10 represents a better diagnostic performance for screening purposes or where a high sensitivity is needed. The low sensitivity of the PHQ-9 algorithm scoring approach could be due to rating scale categories, its higher specificity or other factors that warrant further research.

Competing Interests

No competing interests are declared by authors.

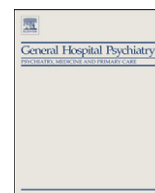
Appendix A. Supplementary data

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.genhosppsych.2014.09.009>.

References

- Gensichen J, Von Korff M, Peitz M, Muth C, Beyer M, Guthlin C, et al. Case management for depression by health care assistants in small primary care practices: A cluster randomized trial. *Ann Intern Med* 2009;151(6):369–78.
- US Preventive Services Task Force. *Guide to Clinical Preventive Services*. 2nd ed. Alexandria, VA: International Medical Publishing; 1996.
- National Institute for Clinical Excellence. *Depression: The treatment and management of depression in adults* (updated edition). London: National Institute for Clinical Excellence; 2009.
- National Institute for Clinical Excellence. *Depression in adults with a chronic physical health problem*. London: National Institute for Clinical Excellence; 2009.
- Canadian Task Force on Preventive Health Care. *Recommendations on screening for depression in adults*. *Can Med Assoc J* 2013;185:775–82.
- Kroenke K, Spitzer R, Williams J. The PHQ-9: Validity of a brief depression severity measure. *J Gen Intern Med* 2001;16(9):606–13.
- Kroenke K, Spitzer R, Williams J. *Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study*. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA* 1999;282(18):1737–44.
- Deeks J. Evaluations of diagnostic and screening tests. In: Davey Smith G, Egger M, Altman DG, editors. *Systematic Reviews in Health Care*. London: BMJ Books; 2000. p. 248–82.
- Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, et al. Conducting systematic reviews of diagnostic studies: Didactic guidelines. *BMC Med Res Methodol* 2002;2:9.
- Centre for Reviews and Dissemination. *Systematic Reviews: CRD's guidance for undertaking reviews in health care*. York: University of York; 2009.
- Lijmer JG, Bossuyt PMM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21(11):1525–37.
- Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21(11):1559–73.
- Song FJ, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol* 2002;31(1):88–95.
- Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. *Can Med Assoc J* 2012;184(3):E191–6.
- Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155(8):529–36.
- Mann R, Hewitt CE, Gilbody SM. Assessing the quality of diagnostic studies using psychometric instruments: Applying QUADAS. *Soc Psychiatry Psychiatr Epidemiol* 2009;44(4):300–7.
- Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58(10):982–90.
- Thompson SG. Systematic review – why sources of heterogeneity in metaanalysis should be investigated. *Br Med J* 1994;309(6965):1351–5.
- Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *Br Med J* 2003;327(7414):557–60.
- Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med* 2010;8(4):348–53.
- Ayalon L, Goldfracht M, Bech P. “Do you think you suffer from depression?” Reevaluating the use of a single item question for the screening of depression in older primary care patients. *Int J Geriatr Psychiatry* 2010;25(5):497–502.
- Diez-Quevedo C, Rangil T, Sanchez-Planell L, Kroenke K, Spitzer RL. Validation and utility of the patient health questionnaire in diagnosing mental disorders in 1003 general hospital Spanish inpatients. *Psychosom Med* 2001;63(4):679–86.
- Eack SM, Greeno CG, Lee B-J. Limitations of the Patient Health Questionnaire in identifying anxiety and depression in community mental health: Many cases are undetected. *Res Soc Work Pract* 2006;16(6):625–31.
- Fann JR, Bombardier CH, Dikmen S, Esselman P, Warms CA, Pelzer E, et al. Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. *J Head Trauma Rehabil* 2005;20(6):501–11.
- Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, et al. Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in east africa. *Psychiatry Res* 2013;210(2):653–61.
- Gjerdingen D, Crow S, McGovern P, Miner M, Center B. Postpartum depression screening at well-child visits: Validity of a 2-question screen and the PHQ-9. *Ann Fam Med* 2009;7(1):63–70.
- Grafe K, Zipfel S, Herzog W, Lowe B. Screening for psychiatric disorders with the Patient Health Questionnaire (PHQ). Results from the German validation study. *Diagnostica* 2004;50(4):171–81.
- Henkel V, Mergl R, Kohonen R, Allgaier A-K, Moller H-J, Hegerl U. Use of brief depression screening tools in primary care: Consideration of heterogeneity in performance in different patient group. *Gen Hosp Psychiatry* 2004;26(3):190–8.
- Hyphantis T, Kotsis K, Voulgari PV, Tsfetaki N, Creed F, Drosos AA. Diagnostic accuracy, internal consistency, and convergent validity of the Greek version of the patient health questionnaire 9 in diagnosing depression in rheumatologic disorders. *Arthritis Care Res* 2011;63(9):1313–21.
- Inagaki M, Ohtsuki T, Yonemoto N, Kawashima Y, Saitoh A, Oikawa Y, et al. 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. *Gen Hosp Psychiatry* 2013;35(6):592–7.
- Khamesh ME, Baradaran H, Javanbakht A, Mirghorbani M, Yadollahi Z, Malek M. Comparison of the CES-D and PHQ-9 depression scales in people with type 2 diabetes in Tehran, Iran. *BMC Psychiatry* 2011;11:61.
- Lamers F, Jonkers CCM, Bosma H, Penninx BWJH, Knottnerus JA, van Eijk J, 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(7):679–87.
- Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry* 2008;8:46.
- Lowe B, Spitzer RL, Grafe K, Kroenke K, Quenter A, Zipfel S, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disord* 2004;78(2):131–40.
- Muramatsu K, Miyaoka H, Kamijima K, Muramatsu Y, Yoshida M, Otsubo T, et al. The patient health questionnaire, Japanese version: Validity according to the mini-international neuropsychiatric interview-plus. *Psychol Rep* 2007;101(3 Pt 1):952–60.
- Navines R, Castellvi P, Moreno-España J, Gimenez D, Udina M, Canizares S, et al. Depressive and anxiety disorders in chronic hepatitis C patients: Reliability and validity of the Patient Health Questionnaire. *J Affect Disord* 2012;138(3):343–51.
- Persoons P, Luyckx K, Desloovere C, Vandenberghje J, Fischler B. Anxiety and mood disorders in otorhinolaryngology outpatients presenting with dizziness: Validation of the self-administered PRIME-MD Patient Health Questionnaire and epidemiology. *Gen Hosp Psychiatry* 2003;25(5):316–23.
- Picardi A, Adler DA, Abeni D, Chang H, Pasquini P, Rogers WH, et al. Screening for depressive disorders in patients with skin diseases: A comparison of three screeners. *Acta Derm Venereol* 2005;85(5):414–9.
- Stafford L, Berk M, Jackson HJ. 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(5):417–24.
- Thekkumpurath P, Walker J, Butcher I, Hodges L, Kleiboer A, O'Connor M, et al. Screening for major depression in cancer outpatients: The diagnostic accuracy of the 9-item patient health questionnaire. *Cancer* 2011;117(1):218–27.
- 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(12):2014–7.
- Thompson AW, Liu H, Hays RD, Katon WJ, Rausch R, Diaz N, et al. Diagnostic accuracy and agreement across three depression assessment measures for Parkinson's disease. *Parkinsonism Relat Disord* 2011;17(1):40–5.
- Turner A, Hambridge J, White J, Carter G, Clover K, Nelson L, et al. Depression screening in stroke: A comparison of alternative measures with the structured diagnostic interview for the diagnostic and statistical manual of mental disorders, fourth edition (major depressive episode) as criterion standard. *Stroke* 2012;43(4):1000–5.
- van Steenberg-Weijnenburg KM, de Vroeg L, Ploeger RR, Brals JW, Vloedveld MG, Veneman TF, et al. Validation of the PHQ-9 as a screening instrument for depression in diabetes patients in specialized outpatient clinics. *BMC Health Serv Res* 2010;10:235.
- Zuithoff NP, Vergouwe Y, King M, Nazareth I, van Wezep MJ, Moons KGM, et al. The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: Consequences of current thresholds in a cross-sectional study. *BMC Fam Pract* 2010;11:1–7.

- [46] Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *J Clin Epidemiol* 2009;62(10):1006–12.
- [47] Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the patient health questionnaire (PHQ): A diagnostic meta-analysis. *J Gen Intern Med* 2007;22(11):1596–602.
- [48] Wittkamp KA, Naeije L, Schene AH, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: A systematic review. *Gen Hosp Psychiatry* 2007;29(5):388–95.
- [49] Williams RT, Heinemann AW, Bode RK, Wilson CS, Fann JR, Tate DG. Improving measurement properties of the Patient Health Questionnaire-9 with rating scale analysis. *Rehabil Psychol* 2009;54(2):198–203.



Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis



Andrew Stephen Moriarty, M.Res., Simon Gilbody, Ph.D., Dean McMillan, Ph.D., Laura Manea, M.Sc. *

Department of Health Sciences, University of York, York, U.K. YO10 5DD

ARTICLE INFO

Article history:

Received 20 March 2015

Revised 11 June 2015

Accepted 12 June 2015

Keywords:

Patient Health Questionnaire

PHQ

PHQ-9

Major depressive disorder

ABSTRACT

Objective: The Patient Health Questionnaire (PHQ-9) is a widely used screening tool for major depressive disorder (MDD), although there is debate surrounding its diagnostic properties. For the PHQ-9, we aimed to:

1. Establish the diagnostic performance at the standard cutoff point (10).
2. Compare the diagnostic performance at the standard cutoff point in different clinical settings.
3. Assess whether there is selective reporting of cutoff points other than 10.

Methods: We searched three databases – Embase, MEDLINE and PSYCHInfo – and performed a reverse citation search in Web of Science. We selected for inclusion studies of any design that assessed the PHQ-9 in adult populations against recognized gold-standard instruments for the diagnosis of either *Diagnostic and Statistical Manual of Mental Disorders* or *International Classification of Diseases* criteria for major depression. Included studies had to report sufficient information to calculate 2*2 contingency tables. Data extraction and synthesis were performed independently by two researchers. For the included studies, we calculated pooled sensitivity, pooled specificity, positive likelihood, negative likelihood ratio and diagnostic odds ratio for cutoff points 7 to 15.

Results: Thirty-six studies (21,292 patients) met inclusion criteria. Pooled sensitivity for cutoff point 10 was 0.78 [95% confidence interval (CI), 0.70–0.84], and pooled specificity was 0.87 (95% CI, 0.84–0.90). At this cutoff, the PHQ-9 is a better screener in primary care than secondary care settings. No conclusions could be drawn at cutoff points other than 10 due to selective reporting of data.

Conclusions: For MDD, the PHQ-9 has acceptable diagnostic properties at cutoff point 10 in different settings. We recommend that future studies report the full range of cutoff points to allow exploration of optimal cutoff points in different settings.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Major depressive disorder (MDD) has a high prevalence in the general population and is associated with considerable morbidity, as well as a high financial cost to society [1]. The Patient Health Questionnaire (PHQ-9) is a self-report tool for screening and case finding for MDD and is based on the Primary Care Evaluation of Mental Disorders, a diagnostic tool developed in the mid-1990s. It is widely used in both clinical and research settings. An indication of its importance comes from its recommendation as a measurement tool for depressive symptoms by the most recent iteration of the *Diagnostic and Statistical Manual of Mental Disorders (DSM, Fifth Edition)*.

Four systematic reviews and meta-analyses previously evaluated the diagnostic properties of the PHQ-9. One of these [2] evaluated how the instrument performs in primary care settings and compared the algorithm scoring method with the summed score ≥ 10 . A meta-analysis published in 2015 by Manea et al. examined the psychometric properties of the PHQ-9 using the algorithm scoring method and compared this scoring method in different settings with the summed score

method at cutoff point of 10 [3]. Another review conducted by Gilbody et al. (published in 2007) summarized the diagnostic properties of the PHQ-9 in different settings [4]. The authors of this review also attempted to summarize the psychometric properties of the PHQ-9 at alternative cutoff points; however, not enough validation studies were found at the time. This analysis was subsequently carried out by Manea et al. in 2012 [5]. This diagnostic meta-analysis has suggested that the performance of the instrument at cutoff point 10 may be lower than that observed in the original validation study. The authors also suggested that different cutoff points may be required for different settings. It is therefore important to examine the performance of other cutoff points which is one of the aims of this review. The review by Manea et al. also highlighted the possibility that there may be selective reporting of cutoff points and that this may artificially inflate the observed diagnostic performance of the measure, at least for cutoff points other than the standard one, which tends to be reported by all studies.

On the basis of this, the current review has three aims: firstly, to establish the diagnostic performance of the PHQ-9 at the standard cutoff point (given the popularity of the PHQ-9, the number of studies available to assess this has grown rapidly since the previous review); secondly, to compare the diagnostic performance of the PHQ-9 at the standard cutoff point in different clinical settings; thirdly, to assess

* Corresponding author.

E-mail address: laura.manea@york.ac.uk (L. Manea).

whether there is selective reporting of cutoff points for cutoffs other than 10.

2. Materials and methods

2.1. Search strategy

We searched Embase, MEDLine and PSYCHInfo from 1999 (when the PHQ-9 was issued) to September 2013 using the terms “PHQ-9,” “PHQ,” “PHQ\$” and “patient health questionnaire.” We manually searched the reference lists of studies fitting the inclusion criteria and performed a reverse citation search in Web of Science. We contacted authors of unpublished studies and conference abstracts in an attempt to minimize publication bias. The search was performed by two independent reviewers (A.M. and L.M.), and any disagreements were resolved by discussion with a third independent reviewer (D.M.).

2.2. Quality assessment

Quality assessment was performed using the updated tool for Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2), which was designed for evaluating the risk of bias and applicability of primary diagnostic accuracy studies when conducting systematic reviews. It covers the areas of patient selection, index test, reference standard, and flow and timing. This tool was adapted for this review, and quality assessments were carried out by two independent reviewers for all studies included in the review. We used a practice run of applying this to five unrelated primary studies to identify and resolve areas of ambiguity or disagreement.

For all domains, we always asked the question of whether there were any concerns regarding the applicability of the patient group, index test or reference test. There was no concern in any of these domains for any of the included studies, and as such, these results are not included in Section 3.

2.3. Study selection

We used the following inclusion–exclusion criteria:

Population: Adult population. Pediatric/adolescent populations were excluded to limit bias. *Instrument:* We included studies that used the PHQ-9. *Comparison (reference standard):* The accuracy of the PHQ-9 had to be assessed against a recognized gold-standard instrument for the diagnosis of either *DSM* or *International Classification of Disease (ICD)* criteria for MDD. Studies were included if the diagnoses were made using a standardized diagnostic structured interview schedule [e.g., Mini International Neuropsychiatric Interview (MINI), Structured Clinical Interview for DSM Disorders (SCID)]. Unguided clinician diagnoses with no reference to a standard structured diagnostic schedule or comparisons of PHQ-9 with other self-report measures were excluded. Studies were also excluded if the target diagnosis was not MDD (e.g., major depressive episode, any depressive disorder). *Outcome:* Studies had to report sufficient information to calculate a 2*2 contingency table for the algorithm. *Study design:* Any design. *Additional criterion:* We avoided double counting of evidence by ensuring that only one study of those which reported overlapping data sets in different journals was included in the meta-analysis. 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.

2.4. Data synthesis and statistical analysis

We constructed 2×2 tables for each cutoff point reported by studies and computed the sensitivity, specificity, and positive and negative predictive values. Pooled estimates of sensitivity, specificity, positive/

negative likelihood ratios and diagnostic odds ratios (DORs) were calculated using random effects bivariate meta-analysis [7]. Summary receiver operator characteristic curves (sROC) were constructed using the bivariate model to produce a 95% confidence ellipse within ROC space [8]. Each data point in the sROC space represents a separate study, unlike a traditional ROC plot which explores the effect of varying thresholds on sensitivity and specificity in a single study.

Heterogeneity was explored using the I^2 statistic based on DORs [9] which describes the percentage of total variation across studies that is caused by heterogeneity rather than chance. I^2 values of 25% may be considered low; 50%, moderate; and 75%, high. We explored the causes of heterogeneity where there was significant between-study heterogeneity. We identified the studies that lay outside of the 95% confidence ellipse by visually inspecting the sROC plots.

We undertook a meta-regression analysis of logit DOR using a priori potential sources of heterogeneity entered as covariates in the meta-regression model [10]. We investigated the heterogeneity resulting from sample or study design characteristics by exploring the effects of potential predictive variables [11].

Finally, publication and small study bias was examined using Begg funnel plots of log DOR versus the inverse of variance [12,13]. Analyses were conducted using STATA version 12, with the `metandi`, `metabias`, `metareg` and `metafunnel` user-written commands.

3. Results

After removing the duplicates, we screened 4513 records for eligibility. Full text was reviewed for 65 papers that met initial inclusion criteria. Thirty-six of 65 met final-stage inclusion criteria. Study selection is summarized in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart in Fig. 1, and further details about the reasons for exclusion are given in Appendix 1.

3.1. Overview of included studies

Thirty-seven citations (36 independent samples) were eligible for inclusion in the review. The characteristics of these studies are reported in Appendix 2.

Of the 36 samples, half (18 studies) were conducted in countries speaking predominantly English [6,14–30]. The other studies used translated versions: Malay [31,32], Portuguese [33,34], Amharic [35], German [36,37], Greek [38], Persian [39], Chinese [40–42], Dutch [43–45], Thai [46], Spanish [47] and Konkani [48]. The mean age of participants in the studies ranged from 23 [24] to 78 [22]. The majority of studies had a wide age range. Three studies validated the PHQ-9 in female populations [21,32,34] and one in a male-only sample [40].

The setting of the studies varied. Fourteen samples recruited participants from primary care [6,15,19,20,22,31,32,34,41,43–46,48], 16 from secondary care [17,18,23,25–30,33,35,38–40,42,47], 4 from community [14,16,21,24] and 2 from mixed settings (outpatient clinics and family practices) [36,37].

None of the included studies reported all cutoff points. The majority of studies reported diagnostic properties of the PHQ-9 at the cutoff point 10. Table 2 presents the number of studies that reported each cutoff point. The percentage of participants who met diagnostic criteria for MDD according to the “gold-standard” reference interview ranged from 1.5% [40] to 43.2% [39]. Some of the studies have a high prevalence of depression because the study design oversampled those who met criteria for depression or were more likely to meet criteria for depression (e.g., oversampled those scoring above a cutoff point 10 on the PHQ-9).

3.2. Methodological quality of included studies

The results of the quality assessment are presented in Appendix 3. The studies included in the review were of mixed methodological quality. Only three studies were judged to have a low risk of bias overall

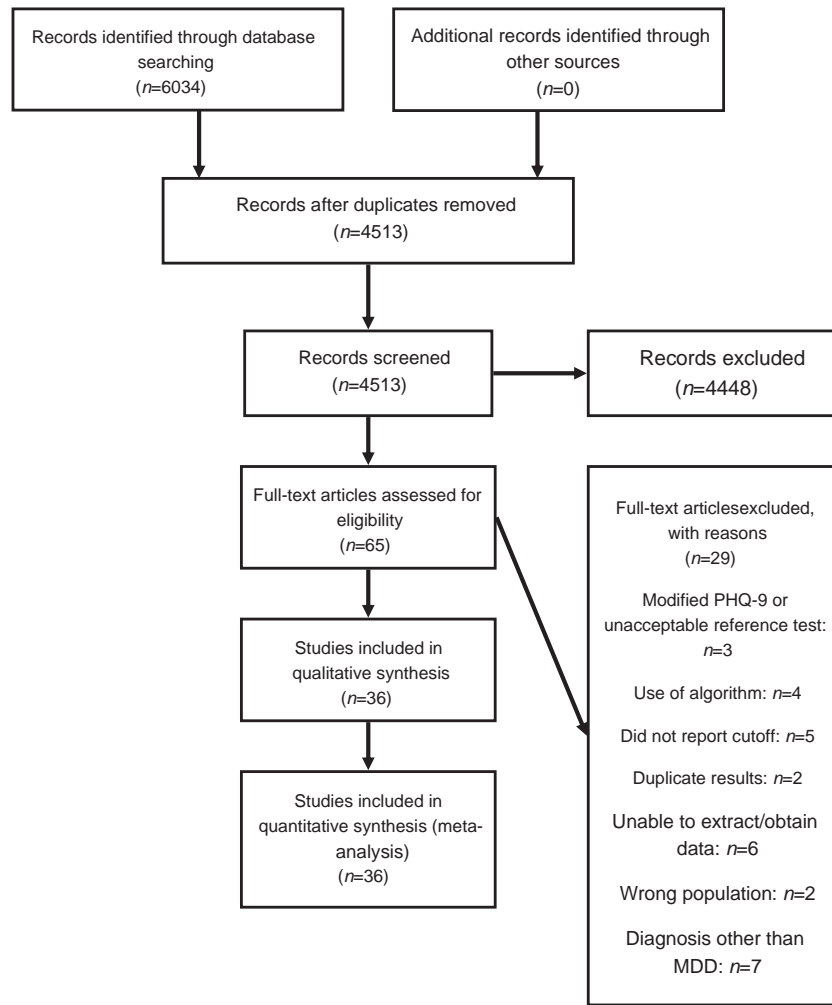


Fig. 1. PRISMA flow diagram outlining study selection.

[6,17,37]. The domain in which most studies were at high risk of bias was “flow and timing” (13 studies). The overall rating of “high risk of bias” in the majority of these studies was a result of failing to report complete data for at least 90% of the patients enrolled in the study. Some of these studies also failed to specify the interval between administration of index test and reference test. Our criterion was that the index test and the reference test must be conducted within 2 weeks of each other for this item to be rated “yes.” This was because fewer than 90% of the participants enrolled in the study were included in the analysis. In the reference test domain, none of the studies was judged to be at high risk of bias.

As the QUADAS-2 ratings indicate, there are a number of additional limitations of the primary studies, and often, details about key methodological criteria were not reported. Blinding in both directions was established in some but not all studies. Lack of blinding may artificially inflate the diagnostic performance of a test. It is possible then that the results may overestimate the performance of the PHQ-9. Many studies were conducted in countries where the native population did not speak English, and in order to be validated, the PHQ-9 needed to be translated in the first instance. Few of these studies though offered details about the translational procedures used.

3.3. Evidence synthesis

3.3.1. Performance of PHQ-9 in detecting MDD at cutoff point 10

Thirty-six studies (21,292 patients: 2573 confirmed cases of MDD by DSM or ICD gold standard) reported diagnostic properties of the PHQ-9

for different cutoff points. Thirty-four out of thirty-six included studies reported the recommended cutoff point of 10. Table 1 shows a summary of this information by cutoff score.

Pooled sensitivity for cutoff point 10 was 0.78 [95% confidence interval (CI), 0.70–0.84], pooled specificity was 0.87 (95% CI, 0.84–0.90), positive likelihood ratio was 6.51 (95% CI, 4.99–8.49), negative likelihood ratio was 0.24 (95% CI, 0.17–0.34), and DOR was 26.27 (95% CI, 16.02–43.07). The heterogeneity was high at 86.0%.

We conducted a meta-regression to explore possible sources of heterogeneity. Descriptive variables and quality assessment criteria (setting, baseline prevalence of MDD, language, whether the study avoided a case–control design and blinding) were examined as predictors. Out of these variables, only the baseline prevalence of MDD was significant ($P=.018$), a higher baseline depression being associated with a higher DOR.

The funnel plot of log DOR versus the inverse of variance (Fig. 2) shows asymmetry that could be due to publication and small study bias. Funnel plots are scatterplots of the treatment effects estimated from individual studies against a measure of study size. In diagnostic test accuracy meta-analyses, the log of odds ratio (OR) and its standard error are represented in funnel plots. The log OR is plotted on the horizontal x-axis, and the standard error of the log OR is plotted on the y-axis. The largest studies have the smallest standard errors, so to place the largest studies at the top of the graph, the vertical axis must be reversed (standard error 0 at the top). In order to aid interpretation of funnel plots, diagonal lines representing the 95% confidence limits around the summary estimate for each standard error on the y-axis are

Table 1

PHQ-9 for diagnosing MDD: heterogeneity and pooled estimates of sensitivity, specificity, positive and negative likelihood ratios, and DORs by cutoff score.

Cutoff point	No of studies	No of patients	Sensitivity (95% CI)	Specificity (95% CI)	Pooled positive likelihood ratio (95% CI)	Pooled negative likelihood ratio (95% CI)	DOR (95% CI)	Heterogeneity: I^2
7	11	8948	0.87 (0.78–0.93)	0.77 (0.69–0.83)	3.88 (2.94–4.13)	0.15 (0.09–0.27)	24.48 (13.77–43.53)	80.8%
8	16	12,501	0.84 (0.76–0.89)	0.82 (0.75–0.87)	4.73 (3.49–6.43)	0.19 (0.13–0.29)	24.62 (13.97–43.38)	78.3%
9	17	11,163	0.84 (0.76–0.89)	0.84 (0.79–0.88)	5.38 (4.05–7.14)	0.18 (0.12–0.28)	28.43 (16.41–49.27)	76.9%
10	34	19,778	0.78 (0.70–0.84)	0.87 (0.84–0.90)	6.51 (4.99–8.49)	0.24 (0.17–0.34)	26.27 (16.02–43.07)	86.0%
11	16	6824	0.83 (0.71–0.90)	0.89 (0.83–0.93)	8.02 (5.11–12.57)	0.18 (0.10–0.32)	43.11 (21.63–85.91)	73.4%
12	16	7762	0.77 (0.67–0.85)	0.91 (0.87–0.94)	9.17 (6.23–13.51)	0.24 (0.16–0.37)	36.16 (22.22–60.33)	66.5%
13	10	2858	0.76 (0.67–0.83)	0.92 (0.86–0.95)	9.87 (5.78–16.87)	0.25 (0.18–0.35)	38.60 (21.68–68.72)	62.0%
14	7	2076	0.62 (0.50–0.74)	0.95 (0.91–0.97)	14.59 (7.89–26.97)	0.38 (0.28–0.53)	37.60 (18.28–77.34)	50.0%
15	10	5531	0.49 (0.38–0.60)	0.96 (0.95–0.97)	14.96 (12.27–18.23)	0.52 (0.42–0.65)	28.56 (19.70–41.41)	50.2%

included. These show the expected distribution in the absence of heterogeneity or of selection biases. Funnel plots were initially proposed as a means of detecting publication bias. However, another possible explanation for funnel plot asymmetry is the exaggeration of the estimate (OR) in small studies of low quality [49]. Fig. 2 suggests that smaller studies showing smaller estimates may be missing. It also suggests that the levels of heterogeneity are high, making the plot asymmetry difficult to interpret. The sROC curve for cutoff 10 is included as Fig. 3.

3.3.2. Performance of PHQ-9 in detecting MDD at cutoff point 10 in different settings

One of the possible reasons for heterogeneity is the various clinical settings in which the PHQ-9 has been validated. On a priori grounds, we conducted subgroup analyses to examine the diagnostic performance of the PHQ-9 in similar clinical settings. The DOR may be used as single indicator of test performance as it is not prevalence dependent. DOR in hospital settings (DOR = 16.51; 95% CI, 8.28–32.91) was lower than that in primary care settings (DOR = 33.41; 95% CI, 14.88–75.01) or community settings (DOR = 40.04; 95% CI, 8.68–184.67). Heterogeneity in all three subgroups was high (primary care I^2 = 85.4%, hospital settings I^2 = 85.8%, community settings I^2 = 88.9%). Table 2 presents a comparative summary of diagnostic properties of the PHQ-9 at cutoff point 10 in three different settings.

3.3.3. Alternative cut points for the PHQ-9

None of the included studies reported the diagnostic properties of the PHQ-9 at all cutoff points (0–27). The average number of cutoff points reported by the included studies was four. Thirty-four studies chose to report cutoff point 10. About half this number of studies

reported the diagnostic accuracy of the PHQ-9 at two cutoff points below or above 10. Hence, the most commonly reported cutoff points were 8, 9, 10, 11 and 12 (Appendix 2). Sensitivity values for these cutoff points showed an unusual trend. While sensitivity values for cutoff points 8, 9 and 11 are almost identical, there is an anomalous drop in sensitivity for cutoff point 10. Sensitivity values for cutoff points 12 and 13 are almost identical as well, and then there is a significant drop from 0.76 to 0.62 for cutoff point 14 and a further big drop for cut-funnel off point 15 from 0.62 to 0.49. The fact that twice the number of studies contributed to calculation of this cutoff point compared to those that reported adjacent cutoff points might explain this unusual result.

Specificity values for cutoff points 8, 9, 10, 11 and 12 followed the expected ascending trend as the cutoff point increases, ranging from 0.82 (95% CI, 0.75–0.87) for cutoff point 8 to 0.91 (95% CI, 0.87–0.94) for cutoff point 12. Heterogeneity was highest for cutoff point 10 (86%).

4. Discussion

4.1. Main findings

Since the first diagnostic meta-analysis of the diagnostic accuracy of the PHQ-9 at different cutoff points was conducted, the number of validation studies that fulfilled the inclusion criteria has doubled. The PHQ-9 has been translated and validated in many languages, countries and settings, and significantly more data were available for this review. However, given that most studies reported a small range of cutoff points, the results for other cutoff points than 10 are more difficult to interpret. If, for instance, 50% of studies in which a cutoff point of 9 showed good psychometric properties are reported but the other 50% of studies are not included in the analysis because, at 9, the psychometric properties were less good, then interpretation of the results is problematic even if the 50% of studies included amount to a big overall number in terms of participants.

The vast majority of studies reported the standard cutoff point, so the findings are unlikely to be substantially affected by the selective reporting of data. For the standard cutoff point, sensitivity is lower than that reported in the original validation study, although specificity is similar. As identified by the previous meta-analysis, at cutoff point 10 the PHQ-9 has a higher sensitivity and similar specificity in primary care as compared to secondary care settings. We pooled the results for the four studies that were conducted in community settings and reported the cutoff point of 10. Pooled sensitivity was 0.76 (95% CI, 0.56–0.81), better than pooled sensitivity in secondary care but less good than pooled sensitivity in primary care; pooled specificity was 0.92 (95% CI, 0.79–0.97), which was better than pooled specificities in primary or secondary care settings. These results however should be interpreted with great caution due to the very high levels of heterogeneity that could not be explained in the meta-regression and were not dealt with by the subgroup analysis.

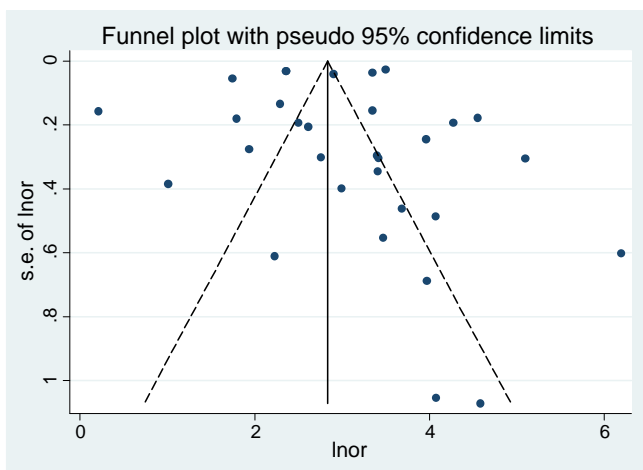


Fig. 2. Funnel plot showing log DOR versus the inverse of variance.

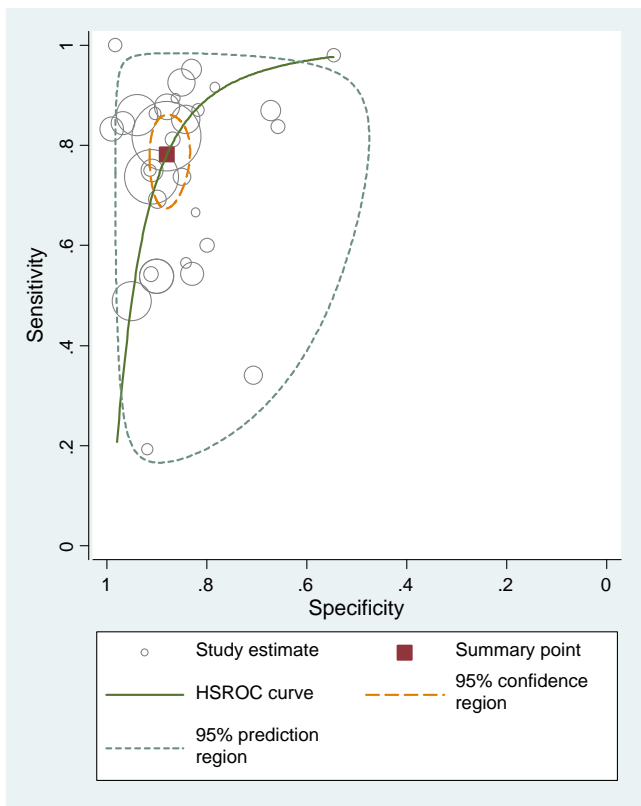


Fig. 3. Summary ROC curve for cutoff point 10.

As described in more detail below, it is difficult to make any firm conclusions about the diagnostic properties of the PHQ-9 at cutoff points other than 10 because different studies chose to report different cutoff points. This is likely to be linked to selective reporting.

4.2. Limitations

The results of the systematic review need to be considered in light of the limitations of the primary studies used in the review and the review itself. The possibility of publication bias could not be ruled out in this review. The funnel plot shows asymmetry that could be due to publication bias; however, given high levels of heterogeneity, the plot is difficult to interpret.

The level of between study heterogeneity was consistently high. A significant finding is that the baseline prevalence of MDD was the only identified predictive source of heterogeneity. This supports one of the hypotheses generated by the previous diagnostic meta-analysis that the same cutoff point may not be appropriate in all settings and that selection of the most appropriate cutoff point should take into account the prevalence of MDD in specific populations. Despite our best efforts to explain the large heterogeneity between studies, we were unable to identify all contributing factors; hence, caution is needed in interpreting the results.

The methodological quality of studies included in the review was mixed. Of particular concern to the current review are studies that

overselected people who were likely to be depressed, which may have introduced partial verification bias [18,29,31,34,35,38,40]. The reported sensitivity and specificity in these studies are likely to be inflated. Although many studies were conducted in countries where the spoken language was not English, very few studies describe the translation process of the PHQ-9 and/or the gold-standard measure used and whether the translations were validated. Poor translation and lack of translation validation can significantly threaten the validity of an instrument.

Perhaps the key limitation of the primary studies is the likelihood of selective reporting of cutoff points other than for the standard one. Sensitivity is expected to fall as a cutoff point increases. However, sensitivity was very similar for cutoff points 8, 9 and 11, but sensitivity for a cutoff point of 10 was unexpectedly low. This suggests that the decision to report the performance of the PHQ-9 at a particular cutoff point may be influenced by whether it performs well in a particular study. Such an approach will capitalize on chance and lead to the artificial inflation of the diagnostic performance of the test at these alternative cutoff points.

5. Conclusions

The aims of the review were to establish the diagnostic performance of the PHQ-9 at the standard cutoff point (10), to compare the diagnostic performance of the PHQ-9 at the standard cutoff point in different clinical settings and to assess whether there is selective reporting of cutoff points other than 10.

Our results further support the conclusions of the previous meta-analysis that the sensitivity of the PHQ-9 at cutoff point 10 is lower than that reported in the original validation study, whereas the specificity is similar. This may have consequences for the use of the PHQ-9 as a screening or case-finding instrument because, often, high sensitivity is required in such circumstances to ensure that few people with depression are missed. One strategy in such circumstances is to lower the cutoff point to increase sensitivity. However, this cannot be recommended currently because it is unclear how the test performs at alternative cutoff points because of selective reporting. A substantial caveat that applies to any recommendations about the performance of the PHQ-9 at the standard cutoff point, however, is that the pooled estimates of sensitivity and specificity were associated with high levels of heterogeneity. One potential explanation of heterogeneity is that the test performs differently in different populations. We examined this through subgroup analyses. There was evidence that the diagnostic properties of the test differed in these populations at the standard cutoff point; however, heterogeneity was also high in each of these subgroup analyses.

We hypothesized in the previous diagnostic meta-analysis of the PHQ-9 that the same single threshold might not be appropriate in all settings. Owing to selective reporting of cutoff points, we were unable to explore whether different cutoff points perform differently in different settings, and no firm conclusions could be drawn. Most studies selectively reported some cutoff points but not others without stating the reason for this. We strongly recommend that future validation studies report the full range of cutoff points. We also recommend that journals publishing validation studies of screening or case-finding instruments request that data on all cutoff points are provided and that this is made a condition of publication.

Table 2 PHQ-9 for diagnosing MDD: heterogeneity and pooled estimates of sensitivity, specificity, positive and negative likelihood ratios, and DORs by settings at cutoff point 10.

Settings	No of studies	No of patients	Sensitivity (95% CI)	Specificity (95% CI)	Pooled positive likelihood ratio (95% CI)	Pooled negative likelihood ratio (95% CI)	DOR (95% CI)	Heterogeneity: I ²
Primary care	13	8496	0.81 (0.68–0.89)	0.88 (0.84–0.91)	7.15 (5.00–10.22)	0.21 (0.12–0.36)	33.41 (14.88–75.01)	85.4%
Secondary care	13	8534	0.70 (0.56–0.81)	0.87 (0.82–0.91)	5.62 (3.82–8.27)	0.34 (0.22–0.51)	16.51 (8.28–32.91)	88.9%
Community	4	1794	0.76 (0.62–0.86)	0.92 (0.79–0.97)	10.30 (3.31–32.07)	0.25 (0.14–0.44)	40.04 (8.68–184.67)	85.8%

Appendix 1. Reasons for exclusions

Reason for exclusion	Study
Modified PHQ-9 or unacceptable reference test	De Man-Van Ginkel et al., 2012 Weobong et al., 2009 Yeung et al., 2008
Use of algorithm	Eack et al., 2006 Mazzotti et al., 2003 Muramatsu et al., 2007
Did not report cutoff	Persoons et al., 2003 Ayalon et al., 2010 Henkel et al., 2004 Picardi et al., 2005 Thompson et al., 2011 van Steenberg-Weijnenburg et al., 2010
Duplicate results	Bombardier et al., 2012 Spitzer et al., 1999
Unable to extract/obtain data	Chaudron et al., 2011 Chen et al., 2010 Kwan et al., 2012 Nan et al., 2011 Priyanka et al., 2010 Twist et al., 2012
Wrong population	Allgaier et al., 2012 Richardson et al., 2010
Diagnosis other than MDD	Cassin et al., 2013 Davis et al., 2013 Inagaki et al., 2013 Pence et al., 2012 Richardson et al., 2010 Schrag et al., 2012 Sockalingam et al., 2011

Appendix 2. Characteristics of included studies

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard
1. Adewuya et al. (2006)	Country: Nigeria Setting: community (students) Mean age: 24.8 (15–40) Female: 41.2%	N=512 2.5% MDD	Administration: self-report Language: English Cutoffs: 8 to 12	DSM-IV MINI
2. Arroll et al. (2010)	Country: New Zealand Setting: primary care Mean age: 49 (17–99) Female: 61%	N=2642 6.2% MDD	Administration: not stated Language: English Cutoffs: 8, 10, 12, 15	DSM-IV SCID
3. Azah et al. (2005)	Country: Malaysia Setting: primary care Mean age: 38.7 (18–79) Female: 61.7%	N=180 16.6% MDD	Administration: self-report Language: Malay Cutoffs: 5 to 12	DSM-IV CIDI
4. Chagas et al. (2013)	Country: Brazil Setting: secondary care Mean age: not stated Female: 52.7%	N=84 25.5% MDD	Administration: self-report Language: Brazilian Cutoffs: 7 to 10	DSM-IV SCID
5. Delgadillo et al. (2011)	Country: UK Setting: community Mean age: 35 (23–54) Female: 25.2%	N=103 49% MDD	Administration: self-report Language: English Cutoffs: 12	CIS-R
6. de Lima Osorio et al. (2009)	Country: Brazil Setting: primary care Mean age: unclear Female: 100%	N=177 34% MDD	Administration: research assistants Language: Portuguese Cutoffs: 10 to 15	DSM-IV SCID

(continued)

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard
7. Elderon et al. (2011)	Country: USA Setting: secondary care Mean age: unclear Female: 18%	N=1022 18.3% MDD	Administration: self-report Language: English Cutoffs: 10	C-DIS
8. Fann et al. (2005)	Country: US Setting: trauma hospital (inpatients with traumatic brain injury) Mean age: 42 (S.D.=17.9) Female: 29.1%	N=135 16.3% MDD	Administration: telephone administered Language: English Cutoffs: 10	DSM-IV SCID
9. Fine et al. (2013)	Country: USA Setting: primary care (Ohio Army National Guard) Mean age: 31 (17–60) Female: 12%	N=498 21.5% MDD	Administration: telephone administered Language: English Cutoffs: 10, 15	DSM-IV SCID-I
10. Gelaye et al. (2013)	Country: Ethiopia Setting: general hospital Mean age: 34.9 (S.D.=11.6) Female: 63.1%	N=363 12.6% MDD	Administration: researcher administered Language: Amharic Cutoffs: 9 to 11	DSM-IV SCAN
11. Gilbody et al. (2007)	Country: UK Setting: primary care Mean age: 42.5 (S.D. 13.6) Female: 77%	N=96 37.5% MDD	Administration: not stated Language: English Cutoffs: 9 to 13	DSM-IV SCID
12. Gjerdingen et al. (2009)	Country: USA Setting: community Mean age: 29.3 Female: 100%	N=438 4.6% MDD	Administration: telephone or self-report Language: English Cutoffs: 10	DSM-IV SCID
13. Gräfe et al. (2004)	Country: Germany Setting: psychosomatic walk-in clinics and family practices Mean age: 41.9 (S.D.=13.8) Female: 67.8%	N=528 Depressed: 29.2% psychosomatic patients; 6.16% medical patients	Administration: self-report Language: German Cutoffs: 10 to 14	DSM-IV SCID
14. Hyphantis et al. (2011)	Country: Greece Setting: hospital, rheumatology patients Mean age: 54.2 (S.D.=13.5) Female: 74%	N=213 32.4% MDD	Administration: researcher administered Language: Greek Cutoffs: 4 to 16	DSM-IV MINI
15. Khamseh et al. (2011)	Country: Iran Setting: outpatient diabetic clinic Mean age: 56.1 (S.D.=9.6) Female: 51.8%	N=185 43.2% MDD	Administration: self-report Language: Persian Cutoffs: 10,13	DSM-IV SCID
16. Kroenke et al. (2001)	Country: USA Setting: primary care Mean age: 46 (S.D.=17) Female: 66%	N=580 7.1% MDD	Administration: self-report Language: English Cutoffs: 9 to 15	DSM-IV SCID
17. Lai et al. (2010)	Country: Hong Kong	N=551 1.5% MDD	Administration: self-report	DSM-IV SCID

(continued)

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard
18. Lamers et al. (2008)	Setting: general hospital Mean age: 33.4 (18–59) Female: 0% Country: Netherlands Setting: primary care (elderly) Mean age: 71.4 (S.D.=6.90) Female: 48.2%	N=713 Depressed: 10.7%	Language: Chinese version Cutoffs: 7 to 8 Administration: self-report Language: Dutch Cutoffs: 7 to 8	DSM-IV MINI
19. Liu et al. (2011)	Country: Taiwan Setting: primary care Mean age: not specified Female: 60.9%	N=1532 3.3% MDD	Administration: self-report Language: Chinese version Cutoffs: 9 to 11	SCAN
20. Lotrakul et al. (2008)	Country: Thailand Setting: primary care Mean age: 45.0 (S.D.=14.30) Female: 73.7%	N=279 6.8% MDD	Administration: self-report Language: Thai Cutoffs: 7 to 15	DSM-IV MINI
21. Lowe et al. (2004)	Country: Germany Setting: outpatient clinics and family practices Mean age: 41.7 (S.D.=13.8) Female: 67.1%	N=501 13.2% MDD	Administration: self-report Language: German Cutoffs: 11 to 13	DSM-IV SCID
22. Navinés et al. (2012)	Country: Spain Setting: general hospital (patients with chronic HCV) Mean age: 43.4 (S.D.=10.2) Female: 28.6%	N=500 6.4% MDD	Administration: self-report Language: Spanish Cutoffs: 10	DSM-IV SCID
23. Patel et al. (2008)	Country: India Setting: primary care Mean age: 37.5 (18–83) Female: 56.4%	N=299 4.3% MDD	Administration: face-to-face interview Language: not specified Cutoffs: 7 to 15	CIS-R
24. Phelan et al. (2010)	Country: USA Setting: primary care (elderly) Mean age: 78 (S.D.=7) Female: 62%	N=71 12% MDD	Administration: research assistant Language: English Cutoffs: 8 to 12	DSM-IV SCID
25. Rooney et al. (2013)	Country: UK Setting: secondary care (glioma) Mean age: 54.2 (S.D.=12.3) Female: 42.6%	N=129 13.5% MDD	Administration: self-report Language: English Cutoffs: 8 to 11	DSM-IV SCID
26. Sherina et al. (2012)	Country: Malaysia Setting: primary care Mean age: 30.9 (18–81) Female: 100%	N=146 21.2% MDD	Administration: self-report Language: Malay Cutoffs: 10	CIDI
27. Sidebottom et al. (2012)	Country: USA Setting: community	N=745 3.6% MDD	Administration: interview Language:	DSM-IV SCID

(continued)

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard
28. Stafford et al. (2007)	(prenatal) Mean age: 23 (S.D.=.5) Female: 100% Country: Australia Setting: secondary care (cardiac procedures) Mean age: 64.14 (38–91) Female: 19.2%	N=193 18.1% MDD	Administration: self-report Language: English Cutoffs: 10	DSM-IV MINI
29. Thekkumpurath et al. (2010)	Country: UK Setting: hospital (cancer patients) Mean age: 61 Female: 63%	N=782 6.3% MDD (of the whole sample)	Administration: not stated Language: English Cutoffs: 5 to 10	DSM-IV SCID
30. Thombs et al. (2008)	Country: US Setting: hospital (outpatients with coronary heart disease) Mean age: 67 (S.D.=11) Female: 18%	N=1024 22% MDD	Administration: not stated Language: English Cutoffs: 7 to 10	DSM C-DIS
31. Turner et al. (2012)	Country: Australia Setting: stroke patients Mean age: 66.7 (S.D.=13.1) Female: 47.2%	N=72 18% MDD	Administration: self-administered Language: English Cutoffs: 8 to 9	DSM-IV SCID
32. Watnick et al. (2005)	Country: USA Setting: secondary care (dialysis) Mean age: 63 (S.D.=15) Female: 32.3%	N=62 19% MDD	Administration: self-report Language: English Cutoffs: 10	DSM-IV SCID
33. Williams et al. (2005)	Country: USA Setting: secondary care (poststroke) Mean age: unclear Female: unclear	N=316 33.5% MDD	Administration: unclear Language: English Cutoffs: 10	DSM-IV SCID
34. Wittkamp et al. (2009)	Country: Netherlands Setting: primary care Mean age: 49.8 Female: 66.7%	N=664 12.3% MDD	Administration: self-report Language: not specified Cutoffs: 10 and 15	DSM-IV SCIDI
35. Zhang et al. (2013)	Country: Hong Kong Setting: Secondary care (diabetic outpatients) Mean age: 55.1 (S.D.=9.5) Female: 40.8%	N=99 23.2% MDD	Administration: Self-report Language: Chinese version Cutoffs: 15	DSM-IV MINI
36. Zuithoff et al. (2010)	Country: Netherlands Setting: primary care Age (years): M=51 (S.D.=16.7) Female: 63%	N=1338 Depressed: 13%	Administration: self-report Language: Dutch	DSM-IV CIDI

Appendix 3. Methodological quality of included studies

Key:

✓=fulfills criterion

✗=does not fulfil criterion

?=does not provide enough information to assess

Study	Patient selection: consecutive or random sample	Patient selection: avoid case-control/artificially inflated base rate	Patient selection: avoided inappropriate exclusions	Patient selection: overall risk of bias	Index test: PHQ-9 interpreted blind to reference test	Index test: was a threshold pre-specified?	Index test: if translated, appropriate translation	Index test: if translated, psychometric properties reported	Index test: overall risk of bias
1. Adewuya et al. (2006)	✓	✓	✗	Unclear	✓	✓	n/a	n/a	Low
2. Arroll et al. (2010)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
3. Azah et al. (2005)	✓	✗	?	High	✓	✓	✓	✓	Low
4. Chagas et al. (2013)	✓	✓	✓	Low	✓	✓	✓	✓	Low
5. Delgadillo et al. (2011)	✗	✓	✓	High	✓	✓	n/a	n/a	Low
6. de Lima Osorio et al. (2009)	✓	✗	✓	High	?	✗	n/a	n/a	High
7. Elderon et al. (2011)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
8. Fann et al. (2005)	✓	✗	✓	High	✓	✓	n/a	n/a	Low
9. Fine et al. (2013)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear
10. Gelaye et al. (2013)	?	✗	?	High	✓	✗	✓	?	High
11. Gilbody et al. (2007)	?	✓	?	Unclear	✓	✓	n/a	n/a	Low
12. Gjerdingen et al. (2009)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear
13. Gräfe et al. (2004)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear
14. Hyphantis et al. (2011)	✓	✗	✓	High	✓	✓	?	?	Unclear
15. Khamseh et al. (2011)	✓	✓	?	Unclear	✓	✓	✓	✓	Low
16. Kroenke et al. (2011)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
17. Lai et al. (2010)	?	✗	✓	High	✓	✓	✓	?	Unclear
18. Lamers et al. (2008)	✓	✓	✓	Low	✓	✗	n/a	n/a	High
19. Liu et al. (2011)	✓	✓	?	Unclear	✓	✗	✓	?	High
20. Lotrakul et al. (2008)	✓	✓	?	Unclear	✓	✓	✓	?	Unclear
21. Lowe et al. (2004)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
22. Navinés et al. (2012)	✓	✓	✓	Low	✓	✓	✓	✓	Low
23. Patel et al. (2008)	✓	✓	✓	Low	✓	✓	?	?	Unclear
24. Phelan et al. (2010)	✗	✓	✓	High	✓	✗	n/a	n/a	High
25. Rooney et al. (2013)	✓	✓	✓	Low	?	✗	n/a	n/a	High
26. Sherina et al. (2012)	✓	✓	✗	High	✓	✓	✓	✓	Low
27. Sidebottom et al. (2012)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
28. Stafford et al. (2007)	?	✓	✓	Unclear	✓	✓	n/a	n/a	Low
29. Thekkumpurath et al. (2010)	?	✓	✓	Unclear	?	✗	n/a	n/a	High
30. Thombs et al. (2008)	✗	✓	?	High	✓	?	n/a	n/a	Unclear
31. Turner et al. (2012)	✓	✓	✓	Low	✓	✓	n/a	n/a	Low
32. Watnick et al. (2005)	?	✗	✓	High	✓	✓	n/a	n/a	Low
33. Williams et al. (2005)	✓	✓	✓	Low	?	✓	n/a	n/a	Unclear
34. Wittkampf et al. (2009)	✓	✓	✓	Low	✓	?	n/a	n/a	Unclear
35. Zhang et al. (2013)	✓	✓	?	Unclear	?	✓	?	?	Unclear
36. Zuithoff et al. (2010)	✓	✓	✓	Low	✓	✓	✓	?	Unclear

Study	Reference test: reference test correctly classifies target condition	Reference test: reference test interpreted blind to PHQ-9	Reference test: if translated, appropriate translation	Reference test: if translated, psychometric properties reported	Reference test: overall risk of bias	Flow/timing: interval of 2 weeks or less	Flow/timing: all participants receive same reference test	Flow/timing: all participants included in analysis?	Flow/timing: overall risk of bias
1. Adewuya et al. (2006)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
2. Arroll et al. (2010)	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear
3. Azah et al. (2005)	✓	✓	✓	✓	Low	✓	✓	✗	High
4. Chagas et al. (2013)	✓	✓	?	?	Unclear	✓	✓	✗	High
5. Delgadillo et al. (2011)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
6. de Lima Osorio et al. (2009)	✓	?	n/a	n/a	Unclear	?	✓	✓	Unclear
7. Elderon et al. (2011)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
8. Fann et al. (2005)	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
9. Fine et al. (2013)	✓	?	n/a	n/a	Unclear	?	✓	✓	Unclear
10. Gelaye et al. (2013)	✓	✓	✓	✓	Low	✓	✓	✓	Low
11. Gilbody et al. (2007)	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear
12. Gjerdingen et al. (2009)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
13. Gräfe et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
14. Hyphantis et al. (2011)	✓	✓	?	?	Unclear	✓	✓	✗	High
15. Khamseh et al. (2011)	✓	✓	✓	✓	Low	✓	✓	?	Unclear
16. Kroenke et al. (2011)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
17. Lai et al. (2010)	✓	✓	?	?	Unclear	?	✗	✗	High
18. Lamers et al. (2008)	✓	✓	n/a	n/a	Low	✓	✓	✗	High

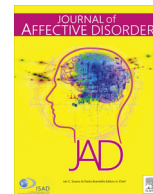
(continued)

Study	Reference test: reference test correctly classifies target condition	Reference test: reference test interpreted blind to PHQ-9	Reference test: if translated, appropriate translation	Reference test: if translated, psychometric properties reported	Reference test: overall risk of bias	Flow/timing: interval of 2 weeks or less	Flow/timing: all participants receive same reference test	Flow/timing: all participants included in analysis?	Flow/timing: overall risk of bias
19. Liu et al. (2011)	✓	✓	✓	✓	Low	✓	✓	?	Unclear
20. Lotrakul et al. (2008)	✓	✓	✓	✓	Low	?	✓	×	High
21. Lowe et al. (2004)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
22. Navinés et al. (2012)	✓	✓	✓	✓	Low	✓	✓	✓	Low
23. Patel et al. (2008)	✓	✓	✓	?	Unclear	?	✓	×	High
24. Phelan et al. (2010)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
25. Rooney et al. (2013)	✓	?	n/a	n/a	Unclear	?	✓	×	High
26. Sherina et al. (2012)	✓	✓	✓	✓	Low	✓	✓	✓	Low
27. Sidebottom et al. (2012)	✓	✓	n/a	n/a	Low	✓	✓	×	High
28. Stafford et al. (2007)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
29. Thekkumpurath et al. (2010)	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear
30. Thombs et al. (2008)	?	✓	n/a	n/a	Unclear	✓	✓	✓	Low
31. Turner et al. (2012)	✓	✓	n/a	n/a	Low	?	✓	✓	Unclear
32. Watnick et al. (2005)	✓	✓	n/a	n/a	Low	✓	✓	✓	Low
33. Williams et al. (2005)	✓	?	n/a	n/a	Unclear	?	✓	✓	Unclear
34. Wittkamp et al. (2009)	✓	✓	n/a	n/a	Low	?	✓	×	High
35. Zhang et al. (2013)	✓	?	✓	✓	Unclear	×	✓	×	High
36. Zuihoff et al. (2010)	✓	✓	?	?	Unclear	?	✓	✓	Unclear

References

- [1] Valenstein M, Vijan S, Zeber JE, Boehm K, Buttar A. The cost–utility of screening for depression in primary care. *Ann Intern Med* 2001;134(5):345–60.
- [2] Wittkamp KA, Naeije L, Schene AH, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review [references]. *Gen Hosp Psychiatry* 2007;29(5):388–95.
- [3] Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *Gen Hosp Psychiatry* 2015;37(1):67–75.
- [4] Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* 2007;22(11):1596–602.
- [5] Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis [references]. *Can Med Assoc J* 2012;184(3):E191–6.
- [6] Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16(9):606–13.
- [7] Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58(10):982–90.
- [8] Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002;21(9):1237–56.
- [9] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *Br Med J* 2003;327(7414):557–60.
- [10] Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21(11):1559–73.
- [11] Lijmer JG, Bossuyt PMM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21(11):1525–37.
- [12] Song FJ, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol* 2002;31(1):88–95.
- [13] 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(9):882–93.
- [14] Adewuyi AO, Ola BAO, 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(1–2):89–93.
- [15] Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med* 2010;8(4):348–53.
- [16] Delgado J, Payne S, Gilbody S, Godfrey C, Gore S, Jessop D, et al. How reliable is depression screening in alcohol and drug users? A validation of brief and ultra-brief questionnaires. *J Affect Disord* 2011;134(1–3):266–71.
- [17] Elderon L, Smolderen KG, Na B, Whooley MA. Accuracy and prognostic value of American Heart Association: recommended depression screening in patients with coronary heart disease: data from the Heart and Soul Study. *Circulation. Cardiovasc Qual Outcomes* 2011;4(5):533–40.
- [18] Fann JR, Bombardier CH, Dikmen S, Esselman P, Warms C, Pelzer E, et al. Validity of the Patient Health Questionnaire-9 in assessing depression following traumatic brain injury. *J Head Trauma Rehabil* 2005;20(6):501–11.
- [19] Fine TH, Contractor AA, Tamburrino M, Elhai JD, Prescott MR, Cohen GH, et al. Validation of the telephone-administered PHQ-9 against the in-person administered SCID-I major depression module. *J Affect Disord* 2013;150:1001–7.
- [20] 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(541):650–2.
- [21] Gjerdingen D, Crow S, McGovern P, Miner M, Center B. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *Ann Fam Med* 2009;7(1):63–70.
- [22] Phelan E, Williams B, Meeker K, Bonn K, Frederick J, LoGerfo J. A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam Pract* 2010;11:63.
- [23] Rooney AG, McNamara S, Mackinnon M, Fraser M, Rampling R, Carson A. Screening for major depressive disorder in adults with cerebral glioma: an initial validation of 3 self-report instruments. *Neuro-Oncology* 2013;15(1):122–9.
- [24] Sidebottom AC, Harrison PA, Godecker A, Kim H. Validation of the Patient Health Questionnaire (PHQ)-9 for prenatal depression screening. *Arch Womens Ment Health* 2012;15(5):367–74.
- [25] Stafford L, Berk M, Jackson HJ. 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(5):417–24.
- [26] Thekkumpurath P, Walker J, Butcher I, Hodges L, Kleiboer A, O'Connor M, et al. Screening for major depression in cancer outpatients: the diagnostic accuracy of the 9-item Patient Health Questionnaire. *Cancer* 2011;117(1):218–27.
- [27] 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(12):2014–7.
- [28] Turner A, Hambridge J, White J, Carter G, Clover K, Nelson L, et al. Depression screening in stroke: a comparison of alternative measures with the Structured Diagnostic Interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (major depressive episode) as criterion standard. *Stroke* 2012;43(4):1000–5.
- [29] Watnick S, Wang P, Demadura T, Ganzini L. Validation of 2 depression screening tools in dialysis patients. *Am J Kidney Dis* 2005;46(5):919–24.
- [30] Williams LS, Brizendine EJ, Plue L, Bakas T, Tu W, Hendrie H, et al. Performance of the PHQ-9 as a screening tool for depression after stroke. *Stroke* 2005;36(3):635–8.
- [31] Azah M, Shah M. Validation of the Malay version brief Patient Health Questionnaire (PHQ-9) among adult attending family medicine clinics. *Int Med J* 2005;12(4):259–63.
- [32] Sherina MS, Arroll B, Goodyear-Smith F. Criterion validity of the PHQ-9 (Malay version) in a primary care clinic in Malaysia. *Med J Malays* 2012;67(3):309–15.
- [33] Chagas MHN, Tumas V, Rodrigues GR, Machado-de-Sousa JP, Filho AS, Hallack JEC, et al. Validation and internal consistency of Patient Health Questionnaire-9 for major depression in Parkinson's disease. *Age Ageing* 2013;42(5):645–9.
- [34] De Lima Osorio F, Mendes AV, Crippa JA, Loureiro SR. Study of the discriminative validity of the PHQ-9 and PHQ-2 in a sample of Brazilian women in the context of primary health care. *Perspect Psychiatr Care* 2009;45(3):216–27.
- [35] Gelaye B, Williams MA, Lemma S, Deyessa N, Bahretibeb Y, Shibre T, et al. Validity of the Patient Health Questionnaire-9 for depression screening and diagnosis in East Africa. *Psychiatry Res* 2013;210(2):653–61.
- [36] Gräfe K, Zipfel S, Herzog W, Löwe B. Screening for psychiatric disorders with the Patient Health Questionnaire (PHQ). Results from the German validation study. *Diagnostica* 2004;50(4):171–81.
- [37] Lowe B, Spitzer RL, Grafe K, Kroenke K, Quenter A, Zipfel S, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disord* 2004;78(2):131–40.
- [38] Hyphantis T, Kotsis K, Voulgari PV, Tsiptaki N, Creed F, Drosos AA. Diagnostic accuracy, internal consistency, and convergent validity of the Greek version of the patient health questionnaire 9 in diagnosing depression in rheumatologic disorders. *Arthritis Care Res* 2011;63(9):1313–21.

- [39] Khamseh ME, Baradaran HR, Javanbakht A, Mirghorbani M, Yadollahi Z, Malek M. Comparison of the CES-D and PHQ-9 depression scales in people with type 2 diabetes in Tehran, Iran. *BMC Psychiatry* 2011;11:61.
- [40] Lai BP, Tang AKL, Lee DTS, Yip ASK, Chung TKH. Detecting postnatal depression in Chinese men: a comparison of three instruments. *Psychiatry Res* 2010;180(2–3):80–5.
- [41] Liu S, Yeh Z, Huang H, Sun F, Tjung J, Hwang L, et al. Validation of Patient Health Questionnaire for depression screening among primary care patients in Taiwan. *Compr Psychiatry* 2011;52(1):96–101.
- [42] Zhang Y, Ting R, Lam M, Lam J, Nan H, Yeung R. Measuring depressive symptoms using the patient health questionnaire-9 in hong kong chinese subjects with type 2 diabetes. *J Affect Disord* 2013;151(2):660–6.
- [43] Lamers F, Jonkers CCM, Bosma H, Penninx BWJH, Knottnerus A, van Eijk JTM. 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(7):679–87.
- [44] Wittkamp K, van Ravesteijn H, Baas K, van de Hoogen H, Schene A, Bindels P, et al. The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care. *Gen Hosp Psychiatry* 2009;31(5):451–9.
- [45] Zuihoff NPA, Vergouwe Y, King M, Nazareth I, van Wezep MJ, Moons KGM, et al. The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a cross-sectional study. *BMC Fam Pract* 2010;11:98.
- [46] Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry* 2008;8:46.
- [47] Navines R, Castellvi P, Moreno-Espana J, Gimenez D, Udina M, Canizares S, et al. Depressive and anxiety disorders in chronic hepatitis C patients: reliability and validity of the Patient Health Questionnaire. *J Affect Disord* 2012;138(3):343–51.
- [48] Patel V, Araya R, Chowdhary N, King M, Kirkwood B, Nayak S, et al. Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires. *Psychol Med* 2008;38(2):221–8.
- [49] Sterne J. In: Sterne J, editor. *Meta-analysis in Stata: an updated collection from the Stata Journal*. Texas USA: Stata Press; 2009.



Research paper

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



Laura Manea^a, Simon Gilbody^a, Catherine Hewitt^b, Alice North^b, Faye Plummer^b,
Rachel Richardson^b, Brett D. Thombs^{a,b}, Bethany Williams^b, Dean McMillan^{a,*}

^a Hull York Medical School and Department of Health Sciences, University of York, United Kingdom

^b Department of Health Sciences, University of York, United Kingdom

ARTICLE INFO

Article history:

Received 22 March 2016

Received in revised form

28 May 2016

Accepted 3 June 2016

Available online 6 June 2016

Keywords:

Major depression

Screening

Diagnostic accuracy

Phq-2

Ultra-brief screening instruments

Diagnostic meta-analysis

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 1529 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.

© 2016 Published by Elsevier B.V.

1. Introduction

Depression is common and disabling, but its management is suboptimal in primary and secondary care (Gilbody et al., 2008). 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 (Gilbody et al., 2008; Thombs et al., 2012). Some national guidelines recommend it in primary care (U.S. Preventive Services Task Force, 2009), whereas others do not (Joffres et al., 2013; Allaby 2010).

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 (Mitchell and Coyne, 2007). One such very brief screening measure for depression is the two-item Patient Health Questionnaire (PHQ-2) (Kroenke et al., 2003), an abbreviated version of the widely used PHQ-9 (Kroenke et al., 2001). 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 (Kroenke et al., 2003). 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 (Kroenke et al., 2001).

* Correspondence to: Hull York Medical School and Department of Health Sciences, ARRC Building, University of York, YO10 5DD, United Kingdom.

E-mail address: dean.mcmillan@york.ac.uk (D. McMillan).

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 (Gilbody et al., 2007). 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.

2. Methods

2.1. 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 A. 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 (Gilbody et al., 2007; Wittkamp et al., 2007; Kroenke et al., 2010; Manea et al., 2012) and a review of ultra-brief screening instruments for depression (Mitchell and Coyne, 2007).

2.2. 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 its 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.

2.3. 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.

2.4. 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) (Whiting et al., 2011). 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 (Mann et al., 2009).

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.

2.5. 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 $\geq 50\%$ to indicate substantial heterogeneity (Centre for Reviews and Dissemination, 2009). 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 (Centre for Reviews and Dissemination, 2009).

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.

3. 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 (Kroenke et al., 2003; Arroll et al., 2010; Chagas et al., 2011; de Lima Osorio et al., 2009, Osorio et al., 2012; de Man-van Ginkel et al., 2012; Delgadillo et al., 2011; Fiest et al., 2014; Inagaki et al., 2013; Lowe et al., 2005; Margrove et al., 2011a; Phelan et al., 2010a; Richardson et al., 2010a, 2010b; Smith et al., 2010; Thombs et al., 2008a; Tsai et al., 2014; Williams et al., 2005; Zhang et al., 2013; Zuithoff et al., 2010a).

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 (Moher et al., 2009) in Fig. 1 and further details about the reasons for exclusion are given in Appendix B.

3.1. Overview of included studies

Table 1 summarises the characteristics of the included studies. Three studies used general primary care samples (Kroenke et al., 2003; Arroll et al., 2010; Zuithoff et al., 2010b), with a further one focused on older adults in primary care (Phelan et al., 2010b). One study focused on patients with epilepsy, but recruited these from

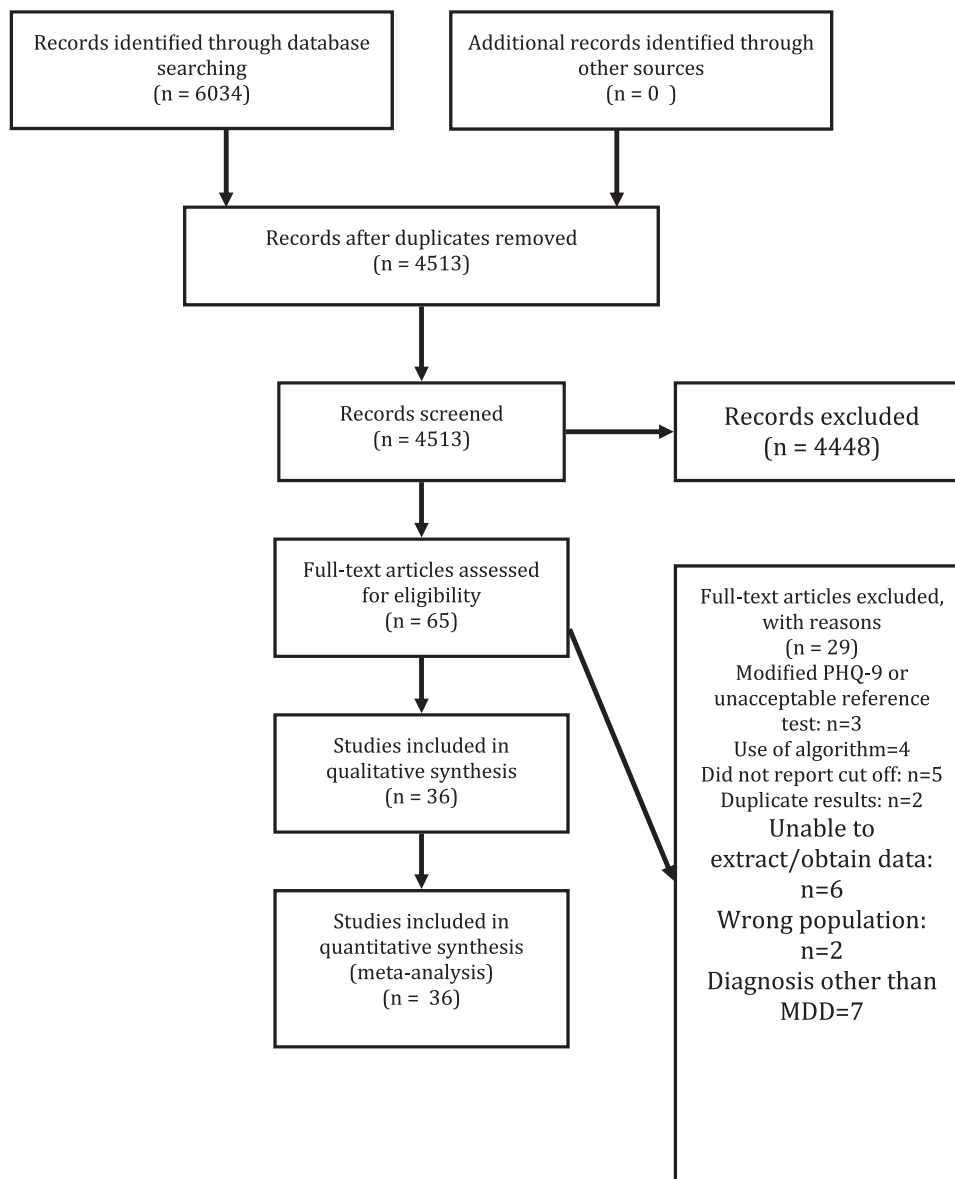


Fig. 1. PRISMA Flow diagram outlining study selection.

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) Female: 39%	N = 100 Depressed: 2%	Administration: Not stated Language: Brazilian Portuguese	DSM-IV SCID CIDI
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)	
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) Female: 59.3%	N = 104 Depressed: 7.4%	Administration: Face to face Language: Japanese	MINI
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) Female: 49.8%	N = 52 Depressed: 48.1%	Administration: Self-report Language: English	DSM-IV SCID
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 Age (yrs): Depressed: 29.31 (sd = 5.98) Non depressed: 28.87 (sd = 6.72) Female: 100%	N = 213 Depressed: 6.1%	Administration: Not stated Language: English	DSM-IV CIDI
Thombs et al. (2008)	Country: US	N = 1024	Administration: Not stated	DSM C-DIS

Table 1 (continued)

Study	Sample characteristics (Country, setting, age, sex)	Sample size and % depressed	PHQ-2 characteristics	Diagnostic standard
	Setting: Outpatients with coronary heart disease Age (yrs): M = 67 (sd = 11) Female: 18%	Depressed: 22%	Language: English	
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 Setting: Community (university students) Age (yrs): M = 21.45 (sd = 1.04) Female: 54.3%	N = 959 Depressed: 8.8%	Administration: Face to face Language: Chinese	DSM-IV SCID
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

primary care (Margrove et al., 2011b). A further three studies used a combination of a primary care setting and another setting, such as outpatient clinics (Lowe et al., 2005; De Lima Osorio et al., 2009; Liu et al., 2011). Eight studies recruited from hospital- or out-patient-based medical specialties (Osorio et al., 2012; de Man-van Ginkel et al., 2012; Fiest et al., 2014; Inagaki et al., 2013; Smith et al., 2010; Williams et al., 2005; Chagas et al., 2011; Thombs et al., 2008b). Of the remainder, one recruited from a community-drug treatment service (Delgado et al., 2011), one from a community-based aging service (Richardson et al., 2010b), one from a research institute focusing on adolescents (Richardson et al., 2010a) and two from community settings (students) (Tsai et al., 2014; Zhang et al., 2013).

All of the studies apart from two (Richardson et al., 2010; Tsai et al., 2014) 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% (Osorio et al., 2012) to 61.2% (Delgado et al., 2011). 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 (Richardson et al., 2010a; Williams et al., 2005; Margrove et al., 2011b).

Six studies stated that a self-report version of the PHQ-2 was used (Kroenke et al., 2003; Delgado et al., 2011; Fiest et al., 2014; Lowe et al., 2005; Tsai et al., 2014; Zuithoff et al., 2010b; Phelan et al., 2010b). In one study it was administered over the telephone (Richardson et al., 2010a) and in four studies it was administered face to face (Chagas et al., 2011; de Man-van Ginkel et al., 2012; Inagaki et al., 2013; Zhang et al., 2013); the remaining studies did not clearly state the method of administration. Translated versions of the PHQ-2 were used in ten studies (Chagas et al., 2011; Osorio et al., 2012; de Man-van Ginkel et al., 2012; Delgado et al., 2011; Inagaki et al., 2013; Lowe et al., 2005; Tsai et al., 2014; Zhang et al., 2013; Zuithoff et al., 2010b; Liu et al., 2011), including Brazilian, Chinese, Dutch, Japanese and German versions.

3.2. 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 (Arroll et al., 2010; Zuithoff et al., 2010b). One of these studies (Zuithoff et al., 2010b), however, was the only one not to meet all of the applicability criteria. The reference standard in Zuithoff et al. (2010b) 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 (Richardson et al., 2010a; Williams et al., 2005; Margrove et al., 2011b).

3.3. 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 .

Nineteen studies reported the performance of the PHQ-2 at cut-off point ≥ 3 . At this cut-off, sensitivity ranged from 0.39 (Thombs et al., 2008a) to 1 (Osorio et al., 2012) and specificity from 0.59 (Smith et al., 2010) to 1 (Margrove et al., 2011b). 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 (Margrove et al., 2011b) 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.

3.4. 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

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
De Lima Osorio et al. (2012)	?	?	✗	High	?	✓	✓	?	Unclear
De Man-van Ginkel et al. (2012)	✓	✓	✓	Low	✓	✓	?	?	Unclear
Delgado 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

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	✓	✓	?	?	Unclear	✓	✓	✗	High

Table 2 (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
et al. (2012)									
Delgado 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)	✓	✗	n/a	n/a	High	✓	✓	✓	Low
(Richardson et al., 2010)									
Richardson et al. (2010)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
(Richardson et al., 2010)									
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

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)	✓	✓	✓
Delgado 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

¹If studies reported multiple cut-off points, 'threshold pre-specified' is coded as not applicable.

Table 3Diagnostic 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–*)
Delgadoillo 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)	* (0.91–*)	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. (2010a)	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. (2010b)	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)

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

* Value could not be estimated.

Table 4Diagnostic 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)

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

* Value could not be estimated.

(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 (Margrove et al., 2011b). A diagnostic meta-analysis was conducted for the remaining four primary care studies (Kroenke et al., 2003; Arroll et al., 2010; Zuithoff et al., 2010b; Phelan et al., 2010b); 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 (Osorio et al., 2012; Inagaki et al., 2013; Smith et al., 2010; Williams et al., 2005; Chagas et al., 2011; Thombs et al., 2008b). 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., 2010b) 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 Fig. 2 for sROC). Heterogeneity was moderate ($I^2=43.5\%$). When the analysis was

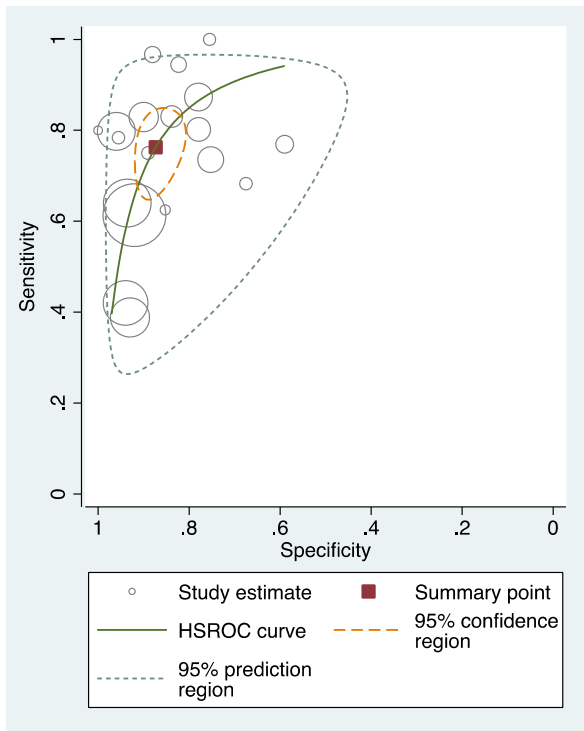


Fig. 2. PHQ-2 at ≥ 3 summary ROC plot of diagnosis of major depressive disorder. Pooled sensitivity and specificity using a bi-variate meta-analysis.

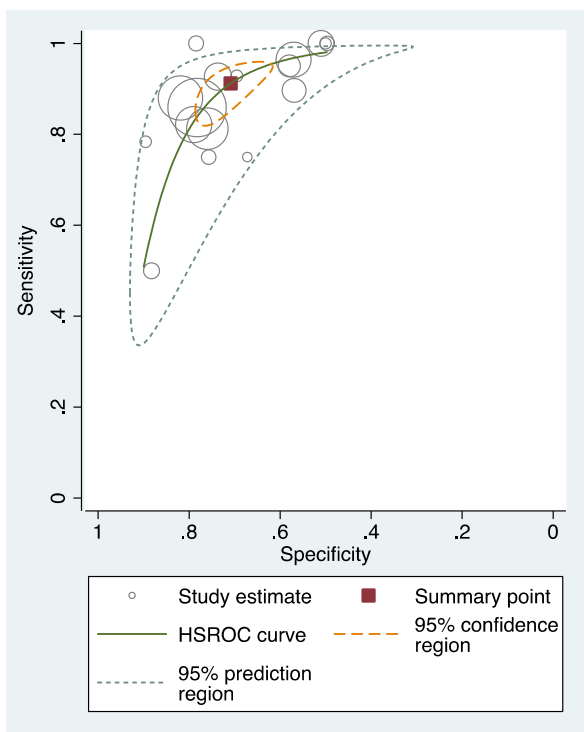


Fig. 3. PHQ-2 at ≥ 2 summary ROC plot of diagnosis of major depressive disorder. Pooled sensitivity and specificity using a bi-variate meta-analysis.

rerun for the four primary care studies (Kroenke et al., 2003; Arroll et al., 2010; Zuihoff et al., 2010b; Phelan et al., 2010b), 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 Fig. 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

(Osorio et al., 2012; de Man-van Ginkel et al., 2012; Fiest et al., 2014; Inagaki et al., 2013; Chagas et al., 2011). 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 ($p > 0.05$).

Fig. 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.

4. 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 (Kroenke et al., 2003). 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 (Kroenke et al., 2003). 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 (Inagaki et al., 2013; De Lima Osorio et al., 2009), all had a lower diagnostic odds ratio than the figure of 43.6 (95% CI = 18.8–101) calculated for Kroenke et al. (2003). 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) (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 (Richardson et al., 2010a; Williams et al., 2005; Margrove et al., 2011b). 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 Fig. 4). For illustrative

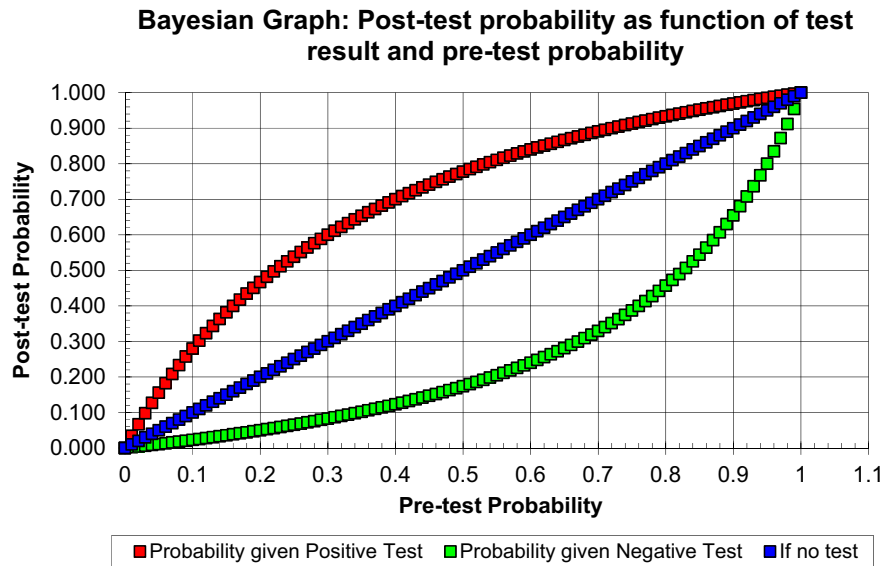


Fig. 4. Performance of PHQ-2 at ≥ 2 using pooled sensitivity and specificity at different prevalence estimates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

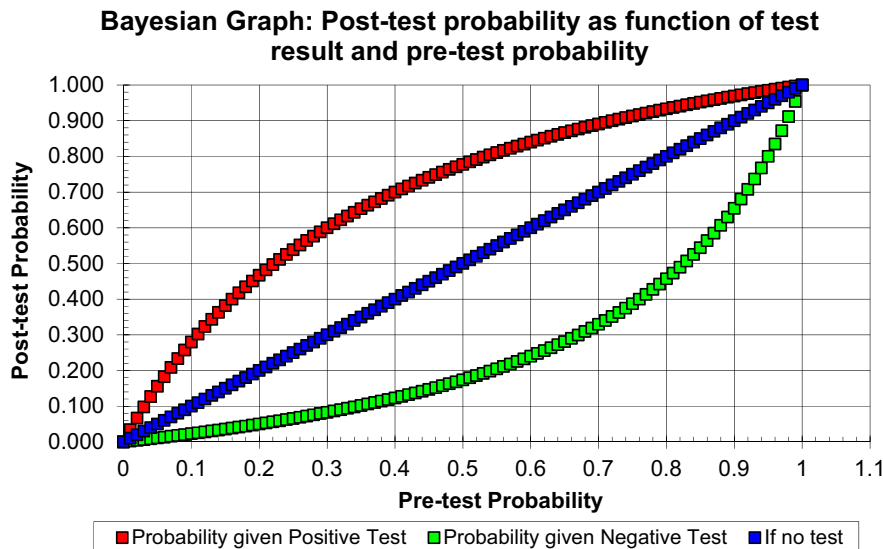


Fig. 5. Performance of PHQ-2 at ≥ 2 using pooled sensitivity and specificity at different prevalence estimates in primary care studies (Gilbody et al., 2007).

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 Fig. 5). This analysis assumes that no patients are being treated for depression, which is perhaps an unrealistic assumption. About half of patients are recognised 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 (Thombs et al., 2011).

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 its 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 (Allaby, 2010).

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-off 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. Fig. 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 Fig. 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 questions appear to be more sensitive but less specific (Bosanquet et al., 2015).

4.1. 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 (Richardson et al., 2010a; Williams et al., 2005; Margrove et al., 2011b) 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 conducted 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 from 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.

4.2. 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) 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.

Conflicts of interest

No authors have any conflicts of interest disclosures.

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. Dr Manea was supported by an NIHR Lectureship award. There was no specific funding for this study, and no funders had any role in the study design, in the collection, analysis or interpretation of data, in the writing of the manuscript or in the decision to submit the manuscript for publication.

Appendix A. 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 B. Excluded studies and reasons for exclusion

see Table B1.

Table B1
 Excluded studies and reasons for exclusion.

Study	Reason for exclusion	Further information
Allgaier et al. (2012)	Reference standard not solely major depression	If either of the two questions were scored as positive, the test was considered positive.
Baker-Glenn et al. (2011)	Non-standard PHQ-2 scoring	
Boyle et al. (2011)	Overlap in sample	Overlap with Richardson et al. (2010a, 2010b)
Brody et al. (n.d.)	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
Thapar 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. (n.d.)	Not PHQ-2	Uses PHQ-9 not PHQ-2
Jiang and Hesser (2011)	Inadequate reference standard	PHQ-8 is treated as the reference standard. (In addition, reference standard is 'any depressive disorder' not major depression.)
Kochhar et al. (2007)	Not PHQ-2	Uses PHQ-9 not PHQ-2 (In addition, reference standard is clinician diagnosis)
Kroenke and 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
Löwe et al. (2005)	Overlap in sample	Overlap with Lowe et al. (2005)
McGuire (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

Table B1 (continued)

Study	Reason for exclusion	Further information
Monahan et al. (2009)	Inadequate reference standard	PHQ-9 used as the reference standard
Pibernik-Okanović et al. (2009)	Reference standard not solely major depression	Reference standard diagnosis combines major depression and dysthymia
Richardson et al.	Overlap in sample	Overlap with Richardson et al. (2010a, 2010b)
Rickels et al. (2009)	Non-standard PHQ-2 scoring	Items are scored yes / no
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 (2011)	Overlap in sample	A review of Richardson et al. (2010a, 2010b)
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.)

References

- Allaby, M., 2010. Screening for Depression: A Report for the UK National Screening Committee (revised report). UK National Screening Committee.
- Allgaier, A.-K., et al., 2012. Screening for depression in adolescents: validity of the patient health questionnaire in pediatric care. *Depress. Anxiety* 29 (10), 906–913 (<http://www.ncbi.nlm.nih.gov/pubmed/22753313>), accessed 25.06.16.
- 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. *Ann. Fam. Med.* 8, 348–353.
- Bosanquet, K., Bailey, D., Gilbody, S., Harden, M., Manea, L., et al., 2015. Diagnostic accuracy of the Whooley questions for the identification of depression: a diagnostic meta-analysis. *BMJ Open*, 5.
- Boyle, L.L., et al., 2011. How do the PHQ-2, the PHQ-9 perform in aging services clients with cognitive impairment? *Int. J. Geriatr. Psychiatry* 26 (9), 952–960 (<http://www.ncbi.nlm.nih.gov/pubmed/21845598>), Accessed June 25, 2016.
- Brody, D.S. et al., Identifying patients with depression in the primary care setting: a more efficient method. *Archiv. Int. Med.*, 158(22), 2469–2475. Available at: (<http://www.ncbi.nlm.nih.gov/pubmed/9855385>) (accessed 25.06.16).
- Bunevicius, A., et al., 2013. Screening for psychological distress in neurosurgical brain tumor patients using the Patient Health Questionnaire-2. *Psycho-oncology* 22 (8), 1895–1900 (<http://www.ncbi.nlm.nih.gov/pubmed/23233453>), accessed 25.06.16.
- Celano, C.M., et al., 2013. Feasibility and utility of screening for depression and anxiety disorders in patients with cardiovascular disease. *Circ. Cardiovas. Qual. Outcomes* 6 (4), 498–504 (<http://www.ncbi.nlm.nih.gov/pubmed/23759474>), accessed 25.06.16.
- Centre for Reviews and Dissemination, 2009. Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care. University Of York, York.
- Chagas, M.H., Crippa, J.A., Loureiro, S.R., Hallak, J.E., 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 Ment. Health* 15, 838–843.
- Chagas, M.H.N., Crippa, J.A.S., Loureiro, S.R., Hallak, J.E.C., 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. *Ment. Health* 15, 838–843.
- Chen, S., et al., 2010. Reliability and validity of the PHQ-9 for screening late-life depression in Chinese primary care. *Int. J. Geriatr. Psychiatry* 25 (11), 1127–1133 (<http://www.ncbi.nlm.nih.gov/pubmed/20029795>), accessed 25.06.16.
- De Lima Osorio, F., Vilela Mendes, A., Crippa, J.A., Loureiro, S.R., 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. *Perspect. Psychiatr. Care* 45, 216–227.
- de Man-van Ginkel, J.M., Gooskens, F., Schepers, V.P., Schuurmans, M.J., Lindeman, E., et al., 2012. Screening for poststroke depression using the patient health questionnaire. *Nurs. Res.* 61 (5), 333–341, Available at: (accessed 10.08.15).
- 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. *J. Affect. Disord.* 134, 266–271.
- Elderon, L., et al., 2011. Accuracy and prognostic value of American Heart Association: recommended depression screening in patients with coronary heart disease: data from the Heart and Soul Study. *Circ. Cardiovas. Qual. Outcomes* 4 (5), 533–540 (<http://www.ncbi.nlm.nih.gov/pubmed/21862720>), accessed 25.06.16.
- Fiest, K.M., Patten, S.B., Wiebe, S., Bullock, A.G.M., Maxwell, C.J., et al., 2014. Validating screening tools for depression in epilepsy. *Epilepsia* 55, 1642–1650.
- 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. *J. Gen. Intern. Med.* 22, 1596–1602.
- Gilbody, S., Sheldon, T., House, A., 2008. Screening and case-finding instruments for depression: a meta-analysis. *Can. Med. Assoc. J.* 178, 997–1003.
- Gjerdengen, D., et al., 2009. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *Ann. Fam. Med.* 7 (1), 63–70 (<http://www.ncbi.nlm.nih.gov/pubmed/19139451>), Accessed June 25, 2016.
- Hahn, D., Reuter, K., Härter, M., 2006. Screening for affective and anxiety disorders in medical patients - comparison of HADS, GHQ-12 and Brief-PHQ. *Psycho-social Med.* 3, Doc09 (<http://www.ncbi.nlm.nih.gov/pubmed/19742274>), accessed 25.06.16.
- Henkel, V. et al., Use of brief depression screening tools in primary care: consideration of heterogeneity in performance in different patient groups. *General hospital psychiatry*, 26(3) 190–198. Available at: (<http://www.ncbi.nlm.nih.gov/pubmed/15121347>) (accessed 25.06.16).
- Henkel, V., et al., 2003. Identifying depression in primary care: a comparison of different methods in a prospective cohort study. *Br. Med. J. (Clinical research ed.)* 326 (7382), 200–201 (<http://www.ncbi.nlm.nih.gov/pubmed/12543837>), accessed 25.06.16.
- Henkel, V., et al., 2004. Screening for depression in primary care: will one or two items suffice? *Eur. Arch. Psychiatry Clin. Neurosci.* 254 (4), 215–223 (<http://www.ncbi.nlm.nih.gov/pubmed/15309389>), accessed 25.06.16.
- 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. *Gen. Hosp. Psychiatry* 35, 592–597.
- Jiang, Y., Hesser, J.E., 2011. A comparison of depression and mental distress indicators, Rhode Island Behavioral Risk Factor Surveillance System, 2006. *Prev. Chron. Dis.* 8 (2), A37 (<http://www.ncbi.nlm.nih.gov/pubmed/21324251>), accessed 25.06.16.
- Joffres, M., Jaramillo, A., Dickinson, J., Lewin, G., Pottie, K., et al., 2013. Recommendations on screening for depression in adults. *Can. Med. Assoc. J. (J. de l'Assoc. Med. Can.)* 185, 775–782.
- Kochhar, P., Rajadhyaksha, S., Suvama, V., 2007. Translation and validation of brief patient health questionnaire against DSM IV as a tool to diagnose major depressive disorder in Indian patients. *J. Postgrad. Med.* 53 (2), 102 (<http://www.jpgmonline.com/text.asp?2007/53/2/102/32209>), accessed 25.06.16.
- Kroenke, K., Spitzer, R.L., 2002. The PHQ-9. *Psychiatr. Ann.* 32 (9), 509–515.
- Kroenke, K., Spitzer, R.L., Williams, J.B.W., 2001. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16, 606–613.
- Kroenke, K., Spitzer, R.L., Williams, J.B., 2003. The patient health questionnaire-2: validity of a two-item depression screener. *Med. Care* 41, 1284–1292.
- Kroenke, K., Spitzer, R.L., Williams, J.B.W., Lowe, B., 2010. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *Gen. Hosp. Psychiatry* 32, 345–359.
- Li, C., et al., 2007. Validity of the patient health questionnaire 2 (PHQ-2) in identifying major depression in older people. *J. Am. Geriatr. Soc.* 55 (4), 596–602 (<http://www.ncbi.nlm.nih.gov/pubmed/17397440>), accessed 25.06.16.
- Liu, S.I., Yeh, Z.T., Huang, H.C., Sun, F.J., Tjung, J.J., et al., 2011. Validation of patient health questionnaire for depression screening among primary care patients in Taiwan. *Compr. Psychiatry* 52, 96–101.
- Lowe, B., Kroenke, K., Grafe, K., 2005. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *J. Psychosom. Res.* 58, 163–171.
- Löwe, B., Kroenke, K., Grafe, K., 2005. Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *J. Psychosom. Res.* 58 (2), 163–171 (<http://www.ncbi.nlm.nih.gov/pubmed/15820844>), accessed 25.06.16.
- Manea, L., Gilbody, S., McMillan, D., 2012. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Can. Med. Assoc. J.* 184, E191–E196.
- Mann, R., Hewitt, C.E., Gilbody, S.M., 2009. Assessing the quality of diagnostic studies using psychometric instruments: applying QUADAS. *Soc. Psychiatry Psychiatr. Epidemiol.* 44, 300–307.
- Margrove, K., Mensah, S., Thapar, A., Kerr, M., 2011a. 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 Behav.* 21, 387–390.
- Margrove, K., Mensah, S., Thapar, A., Kerr, M., 2011b. 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 Behav.* 21, 387–390.
- McGuire, A.W., 2011. Depression Screening by Nurses in Hospitalized Acute Coronary Syndrome Patients.
- McManus, D., Pipkin, S.S., Whooley, M.A., 2005. Screening for depression in patients with coronary heart disease (data from the Heart and Soul Study). *The American journal of cardiology* 96 (8), 1076–1081 (<http://www.ncbi.nlm.nih.gov/pubmed/16214441>), accessed 25.06.16.
- Mitchell, A.J., Coyne, J.C.J., 2007. Do ultra-short screening instruments accurately detect depression in primary care? A pooled analysis and meta-analysis of 22 studies. *Br. J. Gen. Pract.* 57, 144–151.
- Mitchell, A.J., et al., 2008. Acceptability of common screening methods used to detect distress and related mood disorders—preferences of cancer specialists and non-specialists. *Psycho-Oncology* 17 (3), 226–236. <http://dx.doi.org/10.1002/pon.1228>, accessed 25.06.16.
- Mitchell, A.J., et al., 2009. Accuracy of specific symptoms in the diagnosis of major depressive disorder in psychiatric out-patients: data from the MIDAS project. *Psychol. Med.* 39 (07), 1107 (http://www.journals.cambridge.org/abstract_S0033291708004674), accessed 25.06.16.
- Mitchell, A.J., Rao, S., Vaze, A., 2010. Do primary care physicians have particular difficulty identifying late-life depression? A meta-analysis stratified by age. *Psychother. Psychosomat.* 79 (5), 285–294. <http://dx.doi.org/10.1159/000318295>, accessed 25.06.16.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group, 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J. Clin. Epidemiol.* 62, 1006–1012.
- Monahan, P.O., et al., 2009. Validity/reliability of PHQ-9 and PHQ-2 depression scales among adults living with HIV/AIDS in western Kenya. *J. Gen. Int. Med.* 24 (2), 189–197 (<http://www.ncbi.nlm.nih.gov/pubmed/19031037>), accessed 25.06.16.
- Osorio, F., Carvalho, A., Fracalossi, T., Crippa, J., Loureiro, E., 2012. Are two items sufficient to screen for depression within the hospital context. *Int. J. Psychiatry Med.* 44, 141–148.
- Phelan, E., Williams, B., Meeker, K., Bonn, K., Frederick, J., et al., 2010a. A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam. Pract.* 11, 1–9.
- Phelan, E., Williams, B., Meeker, K., Bonn, K., Frederick, J., et al., 2010b. A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam. Pract.* 11, 1–9.
- Pibernik-Okanović, M. et al., 2009. Screening performance of a short versus long version of the Patient Health Questionnaire-Depression in outpatients with diabetes.
- Richardson, L.P., McCauley, E., Grossman, D.C., McCarty, C.A., Richards, J., et al., 2010. Evaluation of the patient health questionnaire-9 item for detecting major depression among adolescents. *Pediatrics* 126, 1117–1123.
- Richardson, T.M., He, H., Podgorski, C., Tu, X., Conwell, Y., 2010. Screening depression aging services clients. *Am. J. Geriatr. Psychiatry* 18, 1116–1123.
- Rickels, M.R., et al., 2009. Assessment of anxiety and depression in primary care: value of a four-item questionnaire. *J. Am. Osteopath. Assoc.* 109 (4) 798–219.
- Robison, J., et al., 2002. Screening for depression in middle-aged and older puerto rican primary care patients. *J. Gerontol. Ser. A, Biol. Sci. Med. Sci.* 57 (5), M308–M314 (<http://www.ncbi.nlm.nih.gov/pubmed/11983725>), accessed 25.06.16.
- Rollman, B.L., et al., 2012. A positive 2-item Patient Health Questionnaire depression screen among hospitalized heart failure patients is associated with elevated 12-month mortality. *J. Card. Fail.* 18 (3), 238–245 (<http://www.ncbi.nlm.nih.gov/pubmed/22385945>), accessed 25.06.16.
- Ryan, D.A., et al., 2012. Sensitivity and specificity of the Distress Thermometer and a two-item depression screen (Patient Health Questionnaire-2) with a "help" question for psychological distress and psychiatric morbidity in patients with advanced cancer. *Psycho-oncology* 21 (12), 1275–1284 (<http://www.ncbi.nlm.nih.gov/pubmed/21919118>), accessed 25.06.16.
- Smith, M.V., Gotman, N., Lin, H., Yonkers, K.A., 2010. Do the PHQ-8 and the PHQ-2 accurately screen for depressive disorders in a sample of pregnant women? *Gen. Hosp. Psychiatry* 32, 544–548.
- Smolderen, K.G., et al., 2011. Real-World Lessons From the Implementation of a Depression Screening Protocol in Acute Myocardial Infarction Patients: Implications for the American Heart Association Depression Screening Advisory. *Circ.: Cardiovas. Qual. Outcomes* 4 (3), 283–292. <http://dx.doi.org/10.1161/CIRCOUTCOMES.110.960013>, accessed 25.06.16.
- Thapar, A., et al., 2014. Detecting recurrent major depressive disorder within primary care rapidly and reliably using short questionnaire measures. *The British J. Gen. Pract.: J. R. Coll. Gen. Pract.* 64 (618), e31–e37 (<http://www.ncbi.nlm.nih.gov/pubmed/24567580>), accessed 25.06.16.
- Thombs, B.D., Ziegelstein, R.C., Whooley, M.A., 2008a. 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.* 23, 2014–2017.
- Thombs, B.D., Ziegelstein, R.C., Whooley, M.A., 2008b. 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.* 23, 2014–2017.
- Thombs, B.D., Arthurs, E., El-Baalbaki, G., Meijer, A., Ziegelstein, R.C., 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. *Br. Med. J.* 343.
- Thombs, B.D., Coyne, J.C., Cuijpers, P., de Jonge, P., Gilbody, S., et al., 2012. Rethinking recommendations for screening for depression in primary care. *Can. Med. Assoc. J.* 184, 413–418.
- Tiffin, P.A., 2011. The Patient Health Questionnaire 2-item is a rapid, sensitive and specific screening tool for identifying adolescents with major depression. *Evid.-Based Ment. Health* 13 (4). <http://dx.doi.org/10.1136/ebmh1110>, accessed 25.06.16.
- Tsai, F.J., Huang, Y.H., Liu, H.C., Huang, K.Y., Liu, S.I., 2014. Patient health questionnaire for school-based depression screening among Chinese adolescents.

- Pediatrics, 133.
- U.S. Preventive Services Task Force, 2009. Screening for depression in adults: US preventive services task force recommendation statement. *Ann. Intern. Med.* 151, 784–792.
- Wagner, L.L., et al., 2013. Screening for depression in community-based radiation oncology settings: Results from RTOG 0841. *ASCO Meeting Abstracts* 31 (15_suppl), 9527.
- Watson, L.C., et al., 2009. Practical depression screening in residential care/assisted living: five methods compared with gold standard diagnoses. *Am. J. Geriatric Psychiatry: Off. J. Am. Assoc. Geriatr. Psychiatry* 17 (7), 556–564 (<http://www.ncbi.nlm.nih.gov/pubmed/19554670>), accessed 25.06.16.
- Whiting, P.F., Rutjes, A.W.S., Westwood, M.E., Mallett, S., Deeks, J.J., et al., 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* 155, 529–536.
- Williams, L.S., Brizendine, E.J., 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.
- Wittkamp, K.A., Naeije, L., Schene, A.H., Husyer, J., van Weert, H.C., 2007. Diagnostic accuracy of the mood module of the patient health questionnaire: a systematic review. *Gen. Hosp. Psychiatry* 29, 388–395.
- 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. *J. Affect. Disord.* 151, 660–666.
- Zuithoff, N.P., Vergouwe, Y., King, M., Nazareth, I., van Wezep, M.J., et al., 2010a. The patient health questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a cross-sectional study. *BMC Fam. Pract.*, 11.
- Zuithoff, N.P., Vergouwe, Y., King, M., Nazareth, I., van Wezep, M.J., et al., 2010b. The patient health questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a cross-sectional study. *BMC Fam. Pract.* 11, 1–7.

BMJ Open Diagnostic accuracy of the Whooley questions for the identification of depression: a diagnostic meta-analysis

Katharine Bosanquet,¹ Della Bailey,¹ Simon Gilbody,^{1,2} Melissa Harden,³ Laura Manea,^{1,2} Sarah Nutbrown,¹ Dean McMillan^{1,2}

To cite: Bosanquet K, Bailey D, Gilbody S, *et al*. Diagnostic accuracy of the Whooley questions for the identification of depression: a diagnostic meta-analysis. *BMJ Open* 2015;5:e008913. doi:10.1136/bmjopen-2015-008913

► Prepublication history and additional material is available. To view please visit the journal (<http://dx.doi.org/10.1136/bmjopen-2015-008913>).

Received 27 May 2015

Accepted 9 October 2015



CrossMark

¹Department of Health Sciences, University of York, York, UK

²Hull York Medical School, University of York, York, UK

³Centre for Reviews and Dissemination, University of York, York, UK

Correspondence to

Katharine Bosanquet; kate.bosanquet@york.ac.uk

ABSTRACT

Objectives: To determine the diagnostic accuracy of the Whooley questions in the identification of depression; and, to examine the effect of an additional 'help' question.

Design: Systematic review with random effects bivariate diagnostic meta-analysis. Search strategies included electronic databases, examination of reference lists, and forward citation searches.

Inclusion criteria: Studies were included that provided sufficient data to calculate the diagnostic accuracy of the Whooley questions against a gold standard diagnosis of major depression.

Data extraction: Descriptive information, methodological quality criteria, and 2×2 contingency tables were extracted.

Results: Ten studies met inclusion criteria. Pooled sensitivity was 0.95 (95% CI 0.88 to 0.97) and pooled specificity was 0.65 (95% CI 0.56 to 0.74). Heterogeneity was low ($I^2=24.1\%$). Primary care subgroup analysis gave broadly similar results. Four of the ten studies provided information on the effect of an additional help question. The addition of this question did not consistently improve specificity while retaining high sensitivity as reported in the original validation study.

Conclusions: The two-item Whooley questions have high sensitivity and modest specificity in the detection of depression. The current evidence for the use of an additional help question is not consistent and there is, as yet, insufficient data to recommend its use for screening or case finding.

Trial registration number: CRD42014009695.

INTRODUCTION

Depression is a highly prevalent condition that affects a substantial proportion of the population, varying from around 1 in 4 women to 1 in 10 men.^{1 2} It leads to impairments in functioning that are as significant as those seen in chronic physical health conditions.³ Although depression is a common condition, it is often hard to detect in primary care and other non-psychiatric

Strengths and limitations of this study

- An original study—the first diagnostic accuracy meta-analysis of the Whooley questions as a screening test for depression.
- Using rigorous methodology—strict inclusion/exclusion and quality assessment criteria—identified 10 studies of sufficient quality for inclusion.
- Substantial variability observed in methodological quality of included studies.
- Inconsistency in how Whooley questions are referred to means further relevant studies may have been missed.

settings. Despite the significance of the problem, there is remarkable uncertainty about the value of screening or case finding for depression. The guidance from different Western countries is contradictory,^{4 5} and from a UK health perspective, recommendations offered by different UK bodies are also inconsistent.^{6–10} The UK National Screening Committee¹¹ concluded that there is insufficient evidence to recommend the adoption of screening for depression and also identified a lack of robust evidence for case finding among populations at elevated risk. In contrast, the National Institute of Health and Care Excellence (NICE) guidance recommends that, in the UK, general practitioners (GPs) consider asking two brief questions to identify potential depression in certain patient groups^{7–9} such as people with long-term conditions and women during the perinatal period; if someone responds positively to either question a more comprehensive assessment is carried out, to determine whether or not an individual is depressed.

NICE guidance recommends considering using the Whooley questions,¹² derived from the original Prime-MD,¹³ to identify potential depression. The Whooley questions consist of two questions asking about low mood and loss of interest or pleasure. In the original

validation study, the questions had a sensitivity of 0.95 (0.89 to 0.98) and specificity of 0.56 (0.52 to 0.61). A subsequent validation study added a third question, which asks whether the person wants help with the difficulties identified.¹⁴ Although NICE endorses the use of the Whooley questions, the guidance recognises that this is based on limited evidence of the diagnostic accuracy of the measure. Perhaps as a consequence of this, practitioners also have doubts about the ability of the questions to detect depression.¹⁵ There is further uncertainty about whether the two or three-item version of the questions should be used, with some NICE guidance recommending the use of the third question,⁹—though recent policy changes have seen this removed¹⁰—while other guidance specifically chose not to adopt this additional question because of a lack of evidence on its effectiveness.⁸

The Whooley questions are at the centre of the UK's approach to the identification of depression, yet at the time the UK guidance was published there was limited evidence on the diagnostic performance of the test. It remains unclear whether a review of the current evidence base would lead to a revision of UK guidance. We conducted a systematic review, therefore, to identify all studies that had examined the diagnostic accuracy of the Whooley questions against a gold standard method of establishing a diagnosis of major depression according to internationally recognised criteria. A further component of the review was to assess the effect of the 'help' question in those studies that included it in the screen.

METHOD

A protocol for the systematic review was developed and published on PROSPERO (registration number: CRD42014009695 <http://www.crd.york.ac.uk/PROSPERO/>). We adhered to Centre for Reviews and Dissemination guidance in the conduct of the review and PRISMA guidelines in the reporting of the review.¹⁶

Data sources and searches

The following databases were searched to identify studies assessing the diagnostic test accuracy of the Whooley questions: MEDLINE, MEDLINE In-Process, PsycINFO, EMBASE, Cumulative Index to Nursing & Allied Health (CINAHL Plus), Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts of Reviews of Effects (DARE), and the Health Technology Assessment (HTA) database. A number of additional sources were searched to identify studies in progress, unpublished research or grey literature: Conference Proceedings Citation Index—Science and Social Science, OAIster, ClinicalTrials.gov, Health Services Research Projects in Progress (HSRProj) and the Trip database.

Searches were conducted from 1994—the year the PRIME-MD was published from where the Whooley

questions were derived—to September 2013. No language restrictions or study design filters were applied to the search strategy. In addition, a forward citation search of the Whooley 1997 paper was carried out in the Web of Science database to identify any further papers on the Whooley questions. We examined the reference lists and conducted a reverse-citation search of all included studies.

A search strategy, consisting of relevant free-text terms and subject headings, was developed in MEDLINE (OvidSP) and then adapted for use in the other databases searched. Online supplementary appendix 1 gives the full search strategy for MEDLINE. Furthermore, we contacted key experts in the field to obtain information about potential unpublished data and for clarification on aspects of their work, which consisted of six authors including Whooley *et al*,¹² Arroll and colleagues.^{14 17}

An update of the searches was conducted in April 2015. No further diagnostic accuracy studies using the Whooley questions were found. However, we did observe changes to policy. NICE had amended guidance on perinatal depression (CG192).¹⁰ It now recommends considering asking the Whooley questions alone rather than with the addition of a help question.

Study selection

Studies were selected using a prepiloted form based on the PICO inclusion criteria in the review protocol. Three reviewers assessed titles and abstracts to identify potentially eligible studies. Any queries were discussed with a second reviewer. Full text was obtained for all articles included after this initial screen. Each of these was assessed using the prepiloted form by two reviewers. At each stage any disagreements were resolved by consensus and where necessary arbitration by further reviewers.

Studies that met the following inclusion criteria were included: *Participants/population*: No restrictions were made in terms of the participants or population. *Instrument*: Studies that used either the two-item or three-item Whooley questions were included. The two-item questions had to use the standard Whooley wording, as outlined in the original article.¹²

1. "During the past month, have you often been bothered by feeling down, depressed, or hopeless?" (yes/no)
2. "During the past month, have you often been bothered by little interest or pleasure in doing things?" (yes/no)¹²

For translated versions, the wording had to be derived from the original. The questions also had to be scored as a dichotomous 'yes'/'no'. For the two-item Whooley questions, only studies that defined a positive screen as 'yes' to one or both of the questions were included. Given inconsistencies in the literature about the precise phrasing of the 'help question', all variations in phrasing were accepted. No restrictions were made in terms of mode of administration (eg, telephone or face-to-face) or the person administering the measure (eg, clinician,

researcher or self-administered). *Comparator (reference standard)*: Studies that use a gold standard diagnostic interview to establish a diagnosis of major depression according to international criteria (Diagnostic and Statistical Manual (DSM) or International Classification of Disease (ICD)) were eligible for inclusion. Studies were excluded if the target diagnosis was not solely major depression (eg, any depressive disorder). No restrictions were made in terms of who administered the gold standard or its mode of administration. *Outcome*: For a study to meet inclusion criteria, it had to report sufficient data to extract 2×2 contingency tables for either the two-item Whooley questions or the two-item questions plus an additional help question. *Study design*: No restrictions were made in the type of study design.

Data extraction and quality assessment

Two reviewers independently extracted the following data to a piloted standardised form: (1) descriptive characteristics of the sample and setting (country, setting, age of sample, gender of sample, sample size, proportion depressed); (2) descriptive characteristics of the Whooley (mode of administration, who administered, language); (3) descriptive characteristics of the gold standard (type of gold standard, whether DSM or ICD diagnoses); (4) quality assessment criteria (see below); and (5) the 2×2 contingency tables for the two-item Whooleys and/or two-item Whooleys plus help question against gold standard diagnosis of major depression. Any disagreements were resolved through consensus or, where necessary, arbitration by a third reviewer. Study authors were contacted to provide additional data or clarification as necessary.

Quality assessment was conducted at the study level and used criteria based on the QUADAS-II.¹⁸ The QUADAS-II 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 developed specific guidance on the coding of the questions in the form of a brief field guide.

We retained all of the risk of bias signalling questions and applicability questions, with the exception of one item (prespecified threshold on the index test). This item was removed because the standard method of scoring the Whooley provides a dichotomous cut-off; there is no ordinal or continuous scale that requires the prespecification of a threshold. For the signalling 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 signalling question 'Was there an appropriate interval between the index test and reference standard?' we defined an appropriate interval as less than 2 weeks in keeping with how this item has been applied in previous diagnostic test accuracy studies of depression.¹⁹

We added two additional questions that were applied to studies using translated versions of the Whooley and reference test. For translations of the reference test, we asked whether appropriate forward and back translation methods were used and whether psychometric properties of the translated version were reported. Similarly, we asked whether appropriate translation methods were used and also applied to any translated version of the Whooley. We also added an additional question to establish whether the studies had used strategies to exclude people already known to a service to have depression. This reflects Thombs *et al's*²⁰ concern that studies which include people already known to be depressed may provide an artificially inflated indication of a test's performance, because the typical aim of a screening or case finding tool is to identify depression in those not already known to be depressed. Studies met this criterion if they used strategies to exclude people already known to be depressed, such as excluding people already known to be using psychotropic medication.

Data synthesis and analysis

We constructed 2×2 contingency tables with true positive, true negative, false positive and false negative results. We performed a bivariate diagnostic meta-analysis to obtain pooled estimates of specificity, sensitivity, likelihood ratios, diagnostic ORs and their associated 95% CIs. The bivariate model is a 2-level model which takes into account the precision by which differences in sensitivity and specificity have been calculated while incorporating and estimating the amount of between-study variability in sensitivity and specificity.²¹ A priori subgroup analyses were conducted on descriptive variables and quality assessment criteria.

Heterogeneity

We measured the between study heterogeneity using the I^2 statistic of the pooled diagnostic OR.²² I^2 describes the percentage of total variation across studies, which is caused by heterogeneity rather than chance. The I^2 has a greater statistical power to detect clinical heterogeneity when fewer studies are available compared to other measures of heterogeneity. I^2 values of 25% may be considered low, 50% moderate and 75% high. We explored the causes of heterogeneity where there was significant between-study heterogeneity by visually inspecting the summary receiver operation characteristic curves and identifying the studies that were outside the 95% confidence ellipse. We also undertook a meta-regression analysis of logit diagnostic OR using a priori potential sources of heterogeneity entered as covariates in the meta-regression model.²³

We investigated the heterogeneity resulting from sample or study design characteristics by exploring the effects of potential predictive variables.²⁴ For the sample we examined the effect of language (translated vs not translated), baseline prevalence of major depressive disorder in the screened population, as a proxy measure of

the spectrum of severity of disorder within the screened population, and study settings (primary care vs general hospital). For study quality, we considered blinding (of the assessor to the results of the Whooley questions as well as the gold standard) and whether the studies avoided a case-control design or an artificially inflated base rate of major depression. If these items were important sources of heterogeneity, then they would be predictive in a meta-regression analysis, and would reduce the level of between-study heterogeneity in the meta-regression model.

Analyses were conducted using STATA V.12, with the metandi, metabias, metareg and metafunnel user-written commands.

RESULTS

The initial search identified 6846 unique citations (10 589 citations before de-duplication). Twenty-two of these citations met initial inclusion criteria and were selected for further screening of the full article (figure 1). Ten of the 22 met final stage inclusion criteria. The reasons for exclusion of the 12 studies are as follows: three used the PHQ-2 not the Whooley,^{25–27} for one study we were unable to establish whether the two-item questionnaire used was the Whooley,²⁸ four did not use a gold standard reference test,^{13 29–31} two did not report data on a diagnosis of major depression alone (eg, outcome was any depression diagnosis)^{32 33} and for two it was not possible to extract information to calculate a 2×2 contingency table.^{34 35}

Overview of included studies

Table 1 summarises the characteristics of the included studies. The studies took place in a variety of countries and settings. The samples included adults and older adults and ranged from predominantly male¹² to entirely female samples.^{36 37} Sample sizes ranged from 89³⁸ to over 1000^{14 39} and the proportion depressed according to the gold standard ranged from 3.3%³⁸ to 34%.⁴⁰ Clinicians administered the Whooley questions in the majority of studies. The language of administration was English in six of the studies; translated versions were used in the remainder. A variety of gold standard measures were used, though the CIDI was used in 4 of the 10 studies.

Quality assessment

Table 2 summarises the results of the quality assessment using QUADAS-II. None of the studies was rated as at low risk of bias across all domains. A rating of an unclear risk of bias was the most common rating across the domains. All studies avoided the use of a case-control design. Only three clearly made attempts to exclude people with a known history of depression. Six of the 10 studies provided evidence of blinding in both directions (ie, Whooley interpreted blind to reference, reference interpreted blind to Whooley). In terms of the

QUADAS-2 applicability criteria, all studies were rated as applicable on all three domains.

Diagnostic properties of the Whooley questions (no help question)

Ten studies reported the diagnostic properties of the Whooley questions. One study⁴¹ reported a significantly lower sensitivity and higher specificity than other studies. In the remaining nine studies, the sensitivity ranged between and 0.90³⁹ and 1.00.^{36–38 42} Specificity values ranged between 0.44^{37 42} and 0.78.¹⁴ Table 3 presents the individual performance of the 10 studies including sensitivity, specificity, likelihood ratios and diagnostic ORs and their corresponding 95% CIs.

The pooled sensitivity was 0.95 (CI 0.88 to 0.97), pooled specificity 0.65 (CI 0.56 to 0.74), pooled positive likelihood ratio 2.78 (CI 2.16 to 3.57), pooled negative likelihood ratio 0.07 (CI 0.03 to 0.16) and diagnostic OR 36.91 (17.52 to 77.76). The level of between-study heterogeneity was low ($I^2=24.1\%$). Figure 2 shows the Whooley questions summary receiver operating characteristic plot of major depression diagnosis. Figure 3 shows the posterior probabilities given positive and negative test results. The figure shows that, at the prevalence rate expected in the general population (less than 20%), the probability of a depressed person with a negative test result is very low; whereas the probability of a depressed person with a positive test result is around 40%.

We conducted a meta-regression to explore possible sources of heterogeneity. Descriptive variables and quality assessment criteria (setting, baseline prevalence of major depression, language, whether the study avoided a case-control design and blinding) were examined as predictors. Out of these variables, only the prevalence of major depression was significant ($p=0.026$).

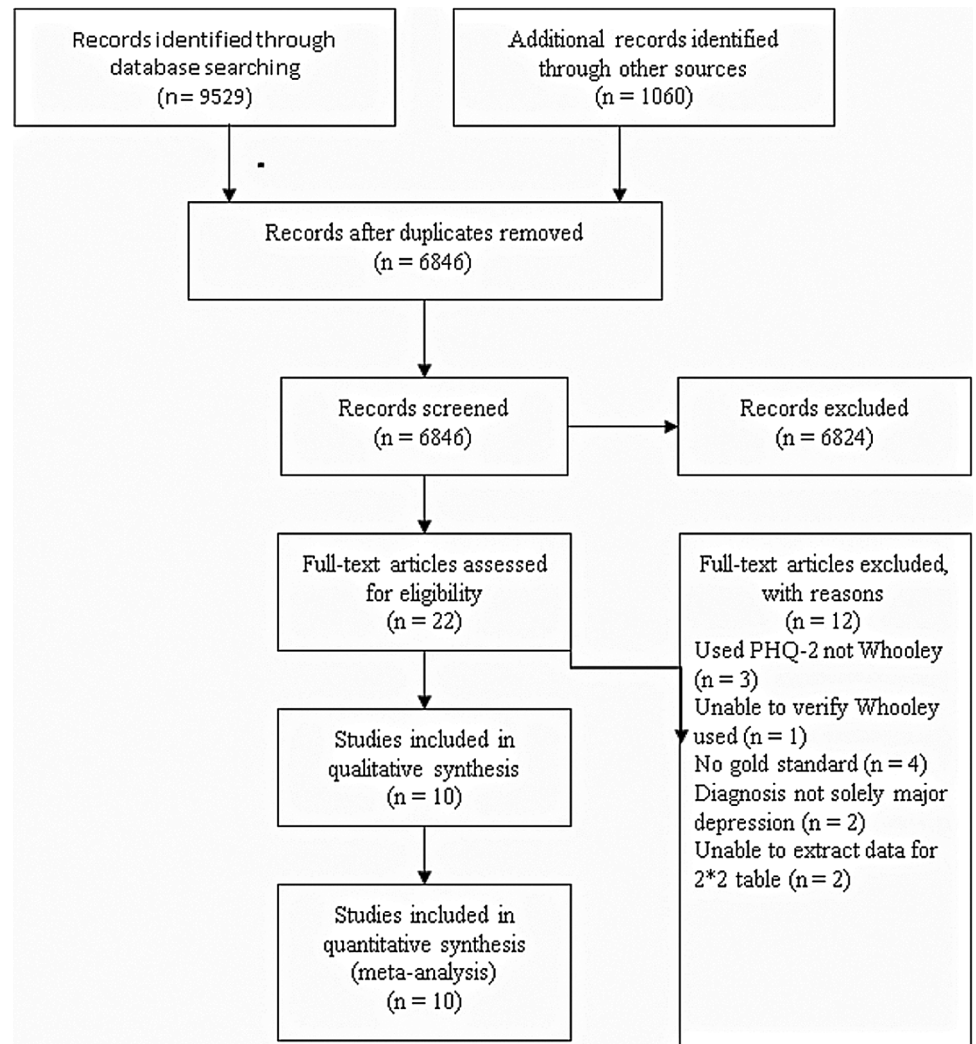
Subgroup analyses

One of the possible reasons for heterogeneity is the various clinical settings in which the Whooley questions have been validated. On a priori grounds we conducted subgroup analyses to examine the diagnostic performance of the Whooley questions in similar clinical settings.

Five studies were conducted in primary care settings,^{14 17 37 40 42} three studies recruited in hospital or out-patient-based medical settings^{12 36 39} and two in community settings.^{38 41} In primary care settings the Whooley questions had a pooled sensitivity of 0.96 (CI 0.91 to 0.98), pooled specificity 0.61 (CI 0.48 to 0.73), pooled positive likelihood ratio 2.53 (CI 1.80 to 3.56), pooled negative likelihood ratio 0.04 (CI 0.01 to 0.13) and diagnostic OR 52.07 (15.65 to 173.18). Heterogeneity in primary care studies was moderate $I^2=49.9\%$.

We did not identify a sufficient number of studies (minimum of four studies for a diagnostic meta-analysis) using a comparable clinical setting to conduct further

Figure 1 Overview of selection of studies (PRISMA).



subgroup analyses for other settings. There were not enough studies to pool the results separately for different age groups.

Six studies validated the original (English) version of the Whooley questions.^{12 14 17 36 37 39} Pooled sensitivity for these studies was 0.95 (0.89 to 0.98), pooled specificity was 0.64 (0.54 to 0.72), positive likelihood ratio 2.67 (2.11 to 3.38), negative likelihood ratio 0.06 (0.02 to 0.15) and pooled diagnostic OR 40.64 (17.00 to 97.14). Heterogeneity in the English studies was low (7.3%).

Whooley questions and help question

Lack of consistency in the phrasing of the questions and how the data were combined meant that we were unable to combine results for a meta-analysis of the help question. Instead we described the results of the studies individually. Two studies^{14 41} considered a positive screen as a positive response to either or both Whooley questions and yes to the help question (yes today; or yes, but not today). The psychometric properties of this method of scoring the Whooley questions were, as reported by Arroll *et al*¹⁴: sensitivity 0.95 (95% CI 0.85 to 0.99), specificity 0.89 (95% CI 0.87 to 0.91), positive likelihood ratio

9.06 (95% CI 7.41 to 11.10) negative likelihood ratio 0.04 (95% CI 0.01 to 0.18) and OR 190.00 95% (50.00—* value unable to be estimated). The psychometric properties reported by Suija *et al* showed a lower sensitivity of 0.68 (95% CI 0.46 to 0.85) but comparable specificity of 0.85 (0.82 to 0.88). Positive likelihood ratio was 4.77 (95% CI 3.36 to 6.78), negative likelihood ratio 0.37 (95% CI 0.21 to 0.66) and OR 12.80 (95% CI 5.40 to 30.20). Arroll *et al*¹⁴ made the distinction between ‘help, yes but not today’ or ‘yes, help today’ though we were unable to extract 2×2 tables for these different responses to the help questions from the data presented in the paper.

The remaining two studies^{36 42} reported the psychometric properties of the help question only in those who scored positive on either Whooley questions. Mann *et al* used the help question ‘is this something you feel you need or want help with?’ rather than the one proposed by Arroll *et al*¹⁴. Psychometric properties of a positive answer to either Whooley question and a positive answer to this question were as follows: sensitivity 0.66 (95% CI 0.38 to 0.88), specificity 0.91 (95% CI 0.78 to 0.98), positive likelihood ratio 8.22 (95% CI 2.62 to 25.80),

Table 1 Descriptive characteristics of the included studies

Study	Sample characteristics (Country, setting, age, sex)	Sample size and % depressed	Whooley characteristics	Diagnostic standard
Adachi <i>et al</i> ³⁸	Country: Japan Setting: community Age (years): M=38.4 (SD=6.6) Female: 9%	N=89 Depressed: 3.3	Administration: psychiatrists and clinical psychologists Language: Japanese	MINI
Arroll <i>et al</i> ¹⁷	Country: New Zealand Setting: primary care Age (years): M=46 (range=16–90) Female: 70%	N=421 Depressed: 6	Administration: general practitioner Language: English	CIDI
Arroll <i>et al</i> ¹⁴	Country: New Zealand Setting: primary care Age (years): not stated Female: % not stated	N=1025 Depressed: 5	Administration: not stated Language: English	CIDI
Gjerdingen <i>et al</i> ³⁷	Country: USA Setting: primary care Age (years): M=28.9 Female: 100%	N=506 Depressed: 4.6	Administration: doctoral-level psychology students Language: English	SCID
Mann <i>et al</i> ³⁶	Country: UK Setting: secondary care Age (years): M=27.4 (SD=5.8) Female: 100%	N=94 Depressed: 19	Administration: Researcher Language: English	SCID
McManus <i>et al</i> ³⁹	Country: USA Setting: secondary care Age (years): M=67 (SD=11) Female: 18%	N=1024 Depressed: 22	Administration: not stated Language: English	DIS
Mohd-Sidik <i>et al</i> ⁴²	Country: Malaysia Setting: primary care Age (years): not stated Female: 100%	N=146 Depressed: 21.2	Administration: family medicine specialist Language: Malay	CIDI
Robison <i>et al</i> ⁴⁰	Country: USA Setting: primary care Age (years): M=61 (range 50–68) Female: 71%	N=303 Depressed: 34	Administration: interviewer Language: Spanish	CIDI
Suija <i>et al</i> ⁴¹	Country: Finland Setting: community Age (years): 72–73 Female: 58.4%	N=474 Depressed: 5.3	Administration: psychiatrist Language: not stated	MINI
Whooley <i>et al</i> ¹²	Country: USA Setting: urgent care clinic Age (years): M=53 (SD=14) Female: 3%	N=536 Depressed: 18.1	Administration: self-report Language: English	DIS

MINI, Mini International Neuropsychiatric Interview; CIDI, Composite International Diagnostic Interview; DIS, Diagnostic Interview Schedule; SCID, Structured Clinical Interview for DSM Disorders; PICO, Population, Intervention, Comparator and Outcome; DOR, Diagnostic Odds Ratio; LR, Likelihood Ratio.

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: appropriately excludes those known to be depressed	Patient selection: overall risk of bias	Index test: Whooley interpreted blind to reference test	Index test: if translated, appropriate translation	Index test: overall risk of bias
		Reference test: Reference test correctly classifies target condition	Reference test: Reference test interpreted blind to Whooley						
Adachi <i>et al</i> ³⁸	✓	✓	?	?	?	Unclear	?	✓	Unclear
Arroll <i>et al</i> ¹⁷	?	✓	✓	✓	✓	Unclear	✓	NA	Low
Arroll <i>et al</i> ¹⁴	✓	✓	✓	✓	✓	Low	✓	NA	Low
Gjerdingen <i>et al</i> ³⁷	✓	✓	×	×	?	High	✓	NA	Low
Mann <i>et al</i> ³⁶	✓	✓	?	?	?	Unclear	✓	NA	Low
McManus <i>et al</i> ³⁹	✓	✓	×	×	?	High	?	NA	Unclear
Mohd Sidik <i>et al</i> (2011)	✓	✓	✓	✓	✓	Low	✓	✓	Unclear
Robison <i>et al</i> ⁴⁰	?	✓	✓	✓	?	Unclear	×	✓	High
Suija <i>et al</i> ⁴¹	✓	✓	✓	✓	×	High	✓	?	Unclear
Whooley <i>et al</i> ¹²	?	✓	✓	✓	?	Unclear	✓	NA	Low

✓, criterion met; ×, criterion not met; ?, insufficient information to code whether criterion met; NA, not applicable.

Table 3 Performance of individual studies (no help question)

Study	Sensitivity (95% CI)	Specificity (95% CI)	Positive LR (95% CI)	Negative LR (95% CI)	DOR (95% CI)
Adachi <i>et al</i> ³⁸	1.00 (0.29 to 1.00)	0.59 (0.48 to 0.69)	2.46 (1.90 to 3.17)	*	*
Arroll <i>et al</i> ¹⁷	0.96 (0.82 to 0.99)	0.67 (0.62 to 0.71)	2.93 (2.51 to 3.43)	0.05 (0.01 to 0.35)	57.10 (9.71 to *)
Arroll <i>et al</i> ¹⁴	0.95 (0.85 to 0.99)	0.78 (0.75 to 0.81)	4.43 (2.86 to 5.09)	0.05 (0.01 to 0.21)	81.70 (21.6 to *)
Gjerdingen <i>et al</i> ³⁷	1.00 (0.92 to 1.00)	0.44 (0.39 to 0.48)	1.79 (1.65 to 1.94)	*	*
Mann <i>et al</i> ³⁶	1.00 (0.78 to 1.00)	0.66 (0.57 to 0.75)	3.00 (2.31 to 3.90)	*	*
McManus <i>et al</i> ³⁹	0.90 (0.85 to 0.93)	0.69 (0.65 to 0.72)	2.91 (2.60 to 3.25)	0.14 (0.09 to 0.21)	20.40 (12.90 to 32.40)
Mohd-Sidik <i>et al</i>	1.00 (0.88 to 1.00)	0.70 (0.61 to 0.78)	3.83 (2.55 to 4.48)	*	*
Robison <i>et al</i> ⁴⁰	0.91 (0.78 to 0.98)	0.44 (0.37 to 0.50)	1.64 (1.42 to 1.89)	0.18 (0.13 to 0.25)	8.90 (2.83 to 27.90)
Suija <i>et al</i> ⁴¹	0.64 (0.42 to 0.82)	0.88 (0.85 to 0.91)	5.75 (3.88 to 8.52)	0.40 (0.24 to 0.68)	14.20 (6.06 to 33.20)
Whooley <i>et al</i> ¹²	0.95 (0.89 to 0.98)	0.56 (0.52 to 0.61)	2.23 (1.98 to 2.50)	0.07 (0.02 to 0.19)	30.80 (11.50 to 81.90)

*Value could not be estimated.

negative likelihood ratio 0.36 (95% CI 0.17 to 0.74) and OR 22.70 (95% CI 4.83 to 105.00).

Mohd-Sidik *et al* used the help question proposed by Arroll *et al*¹⁴, and made the distinction between ‘help, yes but not today’ or ‘yes, help today’. For this study we were able to ascertain how distinguishing between these two options can affect the ability of the help question to detect depression, in people who responded yes to either of the Whooley questions. If a positive answer to the help question was considered ‘yes today’, sensitivity was 0.61 (95% CI 0.42 to 0.78), specificity was 0.94 (95% CI 0.80 to 0.99), positive likelihood ratio was 10.4 (95% CI 2.64 to 41.1), negative likelihood ratio 0.41 (95% CI 0.262 to 0.64) and OR 25.3 (95% CI 5.55—* value unable to be estimated). If a positive answer to help question was considered a positive answer to ‘yes today, or yes, but not today’, sensitivity was higher at 0.87% (95% CI 0.70% to 0.96%), but specificity lower at 0.82% (95% CI 0.65% to 0.93%); positive likelihood ratio was 4.94 (95% CI 2.36 to 10.30), negative likelihood ratio was 0.15 (95% CI 0.06 to 0.39) and OR 31.5 (95% CI 8.22 to 120.00). In this study, therefore, answering ‘yes, help today’ increases the specificity of the Whooley questions when used in conjunction with the help question.

DISCUSSION

NICE guidance recommends that, in the UK, GPs consider using the Whooley questions to identify potential depression in certain patient groups^{7–9} such as people with long-term conditions and women during the perinatal period. The guidance suggests that the Whooley questions are used as a case-finding tool for depression, so if an individual responds positively to one or both of the questions a more comprehensive assessment is carried out to determine whether or not that individual is depressed. The guidance acknowledges, though, that this recommendation is based on limited evidence. Furthermore, there is inconsistency between NICE guidance about whether the Whooley questions should be combined with an additional help question.

This review sought to establish the current evidence for the diagnostic performance of both the original two-item Whooley questions and their combination with an additional help question. The original validation study reported that the two-item version of the questions had high sensitivity (0.95, 95% CI 0.89 to 0.98) and modest specificity (0.56, 95% CI 0.52 to 0.61). The current review found comparable results. Pooled sensitivity was 0.95 (95% CI 0.88 to 0.97) and pooled specificity was 0.65 (95% CI 0.55 to 0.74). Similar figures were also reported in the subgroup analysis examining primary

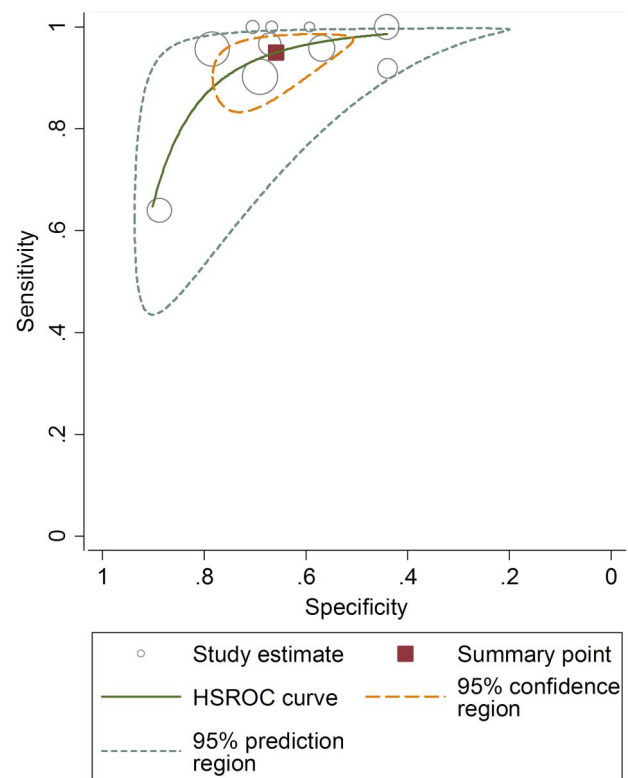
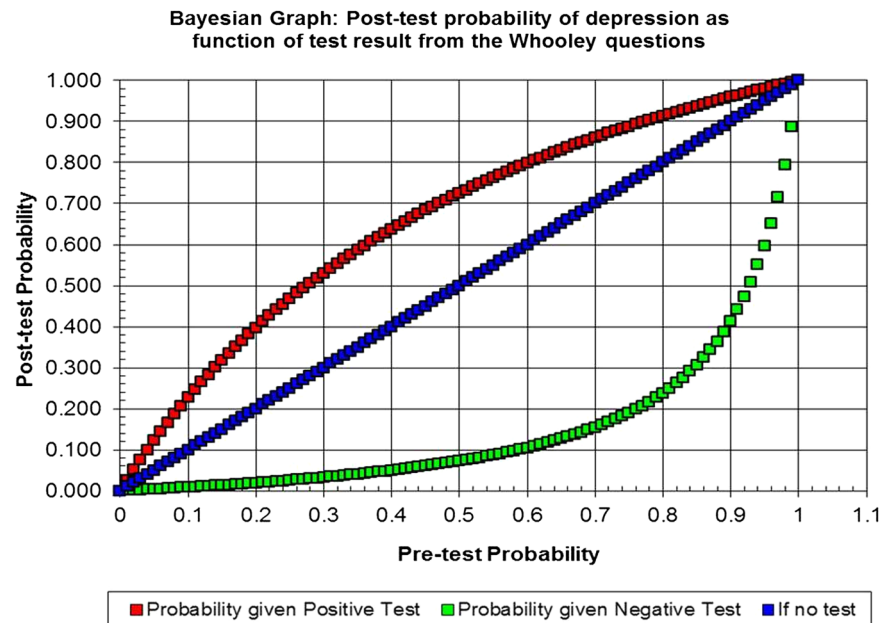


Figure 2 Whooley questions summary receiver operating characteristic plot of diagnosis of major depressive disorder. Pooled sensitivity and specificity using a bivariate meta-analysis.

Figure 3 Bayesian graph for major depressive disorder for Whooley questions.



care studies (sensitivity: 0.96, 95% CI 0.91 to 0.98; specificity: 0.61, 95% CI 0.48 to 0.73).

Our search identified four studies that used the help questions. The authors of the original validation study¹⁴ developed the help question in order to encourage the patient to take an active role in making decisions about their own treatment. They also suggested that the help question may improve specificity. Two categories of help were proposed in this study (help ‘but not today’, and help ‘yes today’).^{14 42} However, of the four studies identified in our review, only two studies, one of which was the original validation study, distinguished between these two help categories: one study combined the two responses⁴¹ and the fourth study³⁶ used a different response. Given the small number of studies and the variability in how the help question was used, we were unable to combine these studies in a meaningful way in order to ascertain the diagnostic performance of the help question when used with the original Whooley questions.

Limitations

The results of the systematic review need to be considered in light of the limitations of the primary studies used in the review and the review itself. As the QUADAS-2 ratings indicate, there are a number of limitations of the primary studies and often details about key methodological criteria were not reported. Only a small number made attempts to exclude people already known to have depression. The aim of depression screening is typically to identify depression in those not known to have that problem. It is possible that excluding those known to be depressed may alter the diagnostic performance of a test. Blinding in both directions was established in some but not all studies. Lack of blinding may artificially inflate the diagnostic performance of a

test. It is possible then that the results may overestimate the performance of the Whooley.

Four of the 10 studies used the CIDI as the reference test, an instrument that has been described as an imperfect gold standard for mental health diagnosis.⁴³ However, the results of these studies for the two-item Whooley questions appeared broadly comparable with studies using a different gold standard. For the studies using the additional help question, the two studies that used the CIDI were the same two studies that reported increased specificity without an impact on sensitivity,^{14 42} findings that were not replicated in the two studies that used other gold standards.^{36 41} It is unclear to what extent these differences are linked to the use of different gold standards.

There are also a number of limitations of the review itself. First, we did not include the ‘help’ question in the search terms, which may have meant we missed articles focused solely on its effect. Second, although efforts were made to identify grey literature, it remains possible that unpublished studies were missed, so we cannot rule out the possibility of publication bias. Third, there is inconsistency in the published studies in how the Whooley questions are referred to, and while the inclusion of various alternative terms for the Whooley questions in the search strategy attempted to address this, it is possible that further relevant studies may have been missed.

Recommendations

The limitations suggest a number of research recommendations. Future diagnostic validation studies should report sufficient detail on the method to permit an assessment of key methodological criteria, such as those given in the QUADAS-2. Subsequent reviews of the Whooley would benefit from a more consistent method

of referring to the Whooley in primary studies. We would recommend the use of the term ‘Whooley questions’ and avoidance of the term ‘PHQ-2’. Although the PHQ-2 shares similarities with the Whooley questions, the PHQ-2⁴⁴ asks about a different time frame and uses a different scoring system (see online supplementary appendix 2). We recommend that future studies should refer to Whooley in the title or abstract to facilitate future reviews of the measure.

CONCLUSION

This review on the diagnostic accuracy of the Whooley questions provides evidence of consistent high sensitivity and moderate specificity for the two questions across a range of settings among different populations. The Whooley questions demonstrate discriminatory power at ruling out depression: few people who answer no to both questions are depressed according to gold standard diagnostic interview. Given that depression is a common condition, this finding should be valuable to clinicians in general practice for use with patients they have concerns about. Despite its modest specificity, which means that many people who score positively will not meet diagnostic criteria for depression, the test retains value in its ability to eliminate the target condition. Although this review identified some evidence that the addition of a help question appeared to improve specificity—when used as second tier test—the inconsistency, both in how the question was phrased and how data were combined, means evidence of its performance remains limited.

Twitter Follow Simon Gilbody at @SimonGilbody

Contributors KB led on all stages of the review from development of the protocol, through screening studies, to data extraction and assessing the quality of the included studies, to production of the final report. DB involved in all stages of the review from development of the protocol, through screening studies and data extraction to synthesis and production of the final report. SG provided expert advice on methodology and approaches to assessment of the evidence base. MH devised the search strategy, carried out the literature searches and wrote the search methodology section of the report. LM reviewed the included studies and assessed their quality, performed the statistical analysis and wrote the results section of the final report. SN involved in the development of the protocol, screening studies for inclusion and data extraction. DM supervised the quality assessment, methodology and approaches to evidence synthesis and provided senior advice and support throughout the review and is guarantor. He contributed to the production of the final report. All parties were involved in drafting and/or commenting on the report.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>

REFERENCES

1. Mental Health Foundation. Mental Health Statistics [cited 2015 07/04/15]. <http://www.mentalhealth.org.uk/help-information/mental-health-statistics/>
2. National Institute for Health and Clinical Excellence. *Clinical knowledge summaries: depression prevalence*. NICE, 2015. [updated Last revised in March 2015; cited 2015 07/04/15]. <http://cks.nice.org.uk/depression/#backgroundsub:1>
3. Moussavi S, Chatterji S, Verdes E, *et al*. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet* 2007;370:851–8.
4. Joffres M, Jaramillo A, Dickinson J, *et al*, Canadian Task Force on Preventive Health Care. Recommendations on screening for depression in adults. *CMAJ* 2013;185:775–82.
5. US Preventive Services Task Force. *Guide to clinical preventive services*. Alexandria, VA: Williams & Wilkins, 1996.
6. Allaby M. *Screening for depression: a report for the National Screening Committee*. Oxford: NHS PHRU, 2010.
7. National Institute for Health and Clinical Excellence. *CG90 depression: the NICE Guideline on the treatment and management of depression in adults*. London, 2010. <http://www.nice.org.uk/guidance/cg90/evidence/cg90-depression-in-adults-full-guidance2>
8. National Institute for Health and Clinical Excellence. *CG91 Depression in adults with a chronic physical health problem*. London, 2010. <http://www.nice.org.uk/guidance/cg91/evidence/cg91-depression-with-a-chronic-physical-health-problem-full-guideline2>
9. National Institute for Health and Clinical Excellence. *Clinical guideline 45: antenatal and postnatal mental health*. London: NICE, 2007.
10. National Institute for Health and Clinical Excellence. *NICE guidelines [CG192]: antenatal and postnatal mental health: clinical management and service guidance*. NICE, 2014. [updated December 2014; cited 2015 08/04/15]. <http://www.nice.org.uk/guidance/cg192/chapter/1-recommendations#recognising-mental-health-problems-in-pregnancy-and-the-postnatal-period-and-referral-2>
11. National Screening Committee. *The UK National Screening Committee's criteria for appraising the viability, effectiveness and appropriateness of a screening programme*. London: NSC, 2003.
12. Whooley M, Avins A, Miranda J, *et al*. Case-finding instruments for depression. Two questions are as good as many. *J Gen Intern Med* 1997;12:439–45.
13. Spitzer R, Williams J, Kroenke K, *et al*. Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. *JAMA* 1994;272:1749–56.
14. Arroll B, Goodyear-Smith F, Kerse N, *et al*. Effect of the addition of a “help” question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study. *BMJ* 2005;331:884.
15. Beauchamp H. What factors influence the use of the Whooley questions by health visitors? *J Health Visiting* 2014;2:378–87.
16. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Int Med* 2009;151:264–9.
17. Arroll B, Khin N, Kerse N. Screening for depression in primary care with two verbally asked questions: cross sectional study. *BMJ* 2003;327:1144–6.
18. Whiting P, Rutjes A, Westwood M, *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Int Med* 2011;155:529–36.
19. Mann R, Hewitt C, Gilbody S. Assessing the quality of diagnostic studies using psychometric instruments: applying QUADAS. *Soc Psychiatry Psychiatr Epidemiol* 2009;44:300–7.
20. Thombs B, Arthurs E, El-Baalbaki G, *et al*. 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. *BMJ* 2011;343:d4825.
21. Reitsma J, Glas A, Rutjes AW, *et al*. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982–90.
22. Higgins J, Thompson S, Deeks J, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.
23. Thompson S, Higgins J. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21:1559–73.
24. Lijmer J, Bossuyt P, Heisterkamp S, *et al*. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21:1525–37.
25. Chagas M, Crippa J, Loureiro S, *et al*. Validity of the PHQ-2 for the screening of major depression in Parkinson's disease: two questions and one important answer. *Aging Ment Health* 2011;15:838–43.

26. Henkel V, Mergl R, Coyne J, *et al.* Screening for depression in primary care: will one or two items suffice? *Eur Arch Psychiatry Clin Neurosci* 2004;254:215–23.
27. Zuithoff N, Vergouwe Y, King M, *et al.* The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a cross-sectional study. *BMC Fam Pract* 2010;11:98.
28. Chochinov HK, Wilson KG, Enns M, *et al.* "Are you depressed?" Screening for depression in the terminally ill. *Am J Psychiatry* 1997;154:674–6.
29. Burton C, Simpson C, Anderson N. Diagnosis and treatment of depression following routine screening in patients with coronary heart disease or diabetes: a database cohort study. *Psychol Med* 2013;43:529–37.
30. Lombardo P, Vaucher P, Haftgoli N, *et al.* The 'help' question doesn't help when screening for major depression: external validation of the three-question screening test for primary care patients managed for physical complaints. *BMC Med* 2011;9:114.
31. Shah M, Karuza J, Rueckmann E, *et al.* Reliability and validity of prehospital case finding for depression and cognitive impairment. *Am Geriatr Soc* 2009;57:697–702.
32. Biswas S, Gupta R, Vanjare H, *et al.* Depression in the elderly in Vellore, South India: the use of a two-question screen. *Int Psychogeriatr* 2009;21:369–71.
33. Ryan D, Gallagher P, Wright S, *et al.* Sensitivity and specificity of the Distress Thermometer and a two-item depression screen (Patient Health Questionnaire-2) with a 'help' question for psychological distress and psychiatric morbidity in patients with advanced cancer. *Psychooncology* 2012;21:1275–84.
34. Brody D, Hahn S, Spitzer R, *et al.* Identifying patients with depression in the primary care setting: a more efficient method. *Arch Intern Med* 1998;158:2469–75.
35. Suzuki T, Nobata R, Kim N, *et al.* Evaluation of Questionnaires (Two question case finding instrument & Beck Depression Inventory) as a tool for screening and intervention of depression in work place. *Seishin Igaku (Clinical Psychiatry)* 2003;45:699–708.
36. Mann R, Adamson J, Gilbody S. Diagnostic accuracy of case-finding questions to identify perinatal depression. *CMAJ* 2012;184:E424–30.
37. 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.
38. Adachi Y, Aleksic B, Nobata R, *et al.* Combination use of Beck Depression Inventory and two-question case-finding instrument as a screening tool for depression in the workplace. *BMJ Open* 2012;2:e000596.
39. McManus D, Pipkin SS, Whooley MA. Screening for depression in patients with coronary heart disease (data from the Heart and Soul Study). *Am J Cardiol* 2005;96:1076–81.
40. Robison J, Gruman C, Gaztambide S, *et al.* Screening for depression in middle-aged and older puerto rican primary care patients. *J Gerontol A Biol Sci Med Sci* 2002;57:M308–14.
41. Suija K, Rajala U, Jokelainen J, *et al.* Validation of the Whooley questions and the Beck Depression Inventory in older adults. *Scand J Prim Health Care* 2012;30:259–64.
42. Mohd-Sidik S, Arroll B, Goodyear-Smith F, *et al.* Screening for depression with a brief questionnaire in a primary care setting: Validation of the two questions with help question (Malay version). *Int J Psychiatry Med* 2011;41:143–54.
43. Gelaye B, Tadesse M, Williams M, *et al.* Assessing validity of a depression screening instrument in the absence of a gold standard. *Ann Epidemiol* 2014;24:527–31.
44. Kroenke K, Spitzer R, Williams J. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Med Care Res Rev* 2003;41:1284–92.

The diagnostic accuracy of brief versions of the Geriatric Depression Scale: a systematic review and meta-analysis

Claire Pocklington¹, Simon Gilbody², Laura Manea² and Dean McMillan²

¹Hull York Medical School, Heslington, York, UK

²Hull York Medical School and Department of Health Sciences, University of York, Heslington, York, UK

Correspondence to: D. McMillan, PhD, E-mail: dean.mcmillan@york.ac.uk

Background: Depression in older adults is often under recognised despite it being the most common mental health illness in this age group. An increasing older adult population highlights the need for improved diagnostic rates. Brief versions (15 items or less) of the Geriatric Depression Scale (GDS), which are suitable for busy clinical practice, could improve detection rates.

Objective: Our aim is to establish the diagnostic accuracy of brief versions of the GDS.

Methods: Twelve electronic databases of published and unpublished literature were searched. Study selection was in accordance with predefined inclusion and exclusion criteria. A recognised gold-standard diagnostic instrument was used as a comparator against data pertaining to the use of a brief version of the GDS in an older adult population. The QUADAS-II was utilised for quality assessment. Narrative analysis and, where possible, meta-analysis were performed.

Results: Thirty-two studies were identified that provided diagnostic data regarding seven brief versions of the GDS (1, 4, 5, 7, 8, 10 and 15-item versions). Pooled sensitivity was 0.89 (95% confidence interval (CI) 0.80–0.94), and specificity was 0.77 (95% CI 0.65–0.86) for the GDS-15 at the recommended cut-off score of 5. Meta-analysis of other brief versions was not possible because of an insufficient number of studies with standardised items.

Conclusions: Results suggest the possibility of selective reporting of cut-off scores, and therefore, findings should be approached cautiously. Studies should report all cut-off scores, and all brief GDS versions should be compiled of standardised items. Copyright © 2016 John Wiley & Sons, Ltd.

Key words: depression; screening; older adults; Geriatric Depression Scale; GDS; meta-analysis

History: Received 5 July 2015; Accepted 25 November 2015; Published online 18 February 2016 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/gps.4407

Introduction

Depression is the commonest mental illness in those aged over 65 years (Anderson, 2001), is associated with higher rates of morbidity and mortality, increased healthcare utilisation and increased economic costs compared with a younger population (Hoyl *et al.*, 1999; Jongenelis *et al.*, 2002; Rinaldi *et al.*, 2003; Nyunt *et al.*, 2009). Despite this, it is often under-recognised and consequently under treated (Birrer and Vemuri, 2004).

One option in such circumstances is to use screening or case finding to identify those with depression. Screening involves applying a screening test to everyone in a particular population. In contrast, case finding is a more

targeted strategy applied to those people known to be at heightened risk of a disorder. For example, the prevalence of certain physical health conditions, such as cardiovascular and cerebrovascular disease, is higher in older adults, and these conditions are also associated with an increased risk of depression (Drayer *et al.*, 2005; Fiske *et al.*, 2009). A case-finding strategy may focus on older adults with these conditions. Clinical assessment will be initiated for all individuals who score positive on a screening test.

The Geriatric Depression Scale (GDS) is a widely used screening tool for depression specifically designed for use in older adults. The measure does not contain somatic symptoms unlike other screening tools for

depression on the basis that these may lack discriminatory capacity in older adults because they may be attributed to comorbid physical conditions and the ageing process (Yesavage *et al.*, 1982). For example, reduced energy levels and appetite, both somatic symptoms of depression, are commonly found secondary to old age and numerous physical health problems (Birrer and Vemuri, 2004).

Although there are existing systematic reviews of the diagnostic accuracy of the GDS (Watson and Pignone, 2003; Wancata *et al.*, 2006; Mitchell *et al.*, 2010a, 2010b; Dennis *et al.*, 2012), these have focused mainly on the full, 30-item version. Time demands in most clinical settings are likely to require briefer tools, but while shorter versions of the GDS exist, these have received less attention. This review focuses, therefore, on versions of the GDS that have 15 or fewer items. The focus on briefer versions reflects current policy and practice recommendations in the UK (NICE, 2011, 2014).

There are additional reasons to conduct a further review. Firstly, the searches for even the most recent review were conducted in 2009 (Dennis *et al.*, 2012). Secondly, there are methodological limitations of the previous reviews. None of the reviews, for example, extensively searched grey literature sources. Four of the five previous reviews did not provide a detailed, standardised quality assessment of the primary studies (Watson and Pignone, 2003; Wancata *et al.*, 2006; Mitchell *et al.*, 2010a; Dennis *et al.*, 2012). It is difficult to draw conclusions about the diagnostic accuracy of the GDS if the methodological limitations of studies contributing to the accuracy estimates are not taken into account. The aim of this review was to provide an up-to-date assessment of the diagnostic accuracy of brief versions of the GDS adhering to best practice guidelines in the conduct and reporting of reviews.

Methods

Literature search

A protocol was developed before commencing the review. The databases MEDLINE, EMBASE, PsycInfo, Cumulative Index to Nursing and Allied Health (CINAHL Plus), Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts and Reviews of Effects (DARE) and the Health Technology Assessment (HTA) database were used to identify studies. Further studies were identified through searching grey literature

and trials registries, which included conference proceedings via Web of Science, ClinicalTrials.gov, British library EThOS, Guideline.gov and OpenGrey.

The search strategy was composed of free text and thesauri terms. It was developed in MEDLINE and then adapted for use in the other databases. Searches were performed from 1982, which is when the GDS was developed, to April 2014. There were no restrictions on language or publication status. Appendix 1 gives the search strategy for MEDLINE.

Study selection

Pre-piloted inclusion criteria were defined in the protocol. A single reviewer (C.P.) reviewed the titles and abstracts to identify eligible studies. Any uncertainty was discussed with a second reviewer (D. McM), and any disagreements were resolved by discussion with a third reviewer (S. G.). For all studies identified as initially eligible, the full text was obtained and reviewed against the inclusion and exclusion criteria.

The following inclusion criteria had to be met. *Population*—study participants 55 years of age or older with no restrictions on setting. *Instrument*—data pertaining to use of the GDS-15 or a briefer version in order to diagnose major depression. A briefer item version of the GDS may have been extracted from the original GDS or the GDS-15. There were no restrictions in terms of administration mode or who administered the GDS to the participants. There were no restrictions in terms of language of the GDS. *Comparator*—the presence of major depression diagnosed using a gold-standard diagnostic interview or instrument that utilised the International Classification of Disease (ICD) or the Diagnostic and Statistical Manual of Mental Disorders (DSM) diagnostic criteria. *Outcome*—eligible studies had to have reported sufficient data to extract a 2×2 contingency table. The 2×2 contingency tables were extracted for any cut-off point for any brief version of the GDS. Study design: no restrictions were made in terms of study design.

Data extraction

Data were entered into a standardised proforma by one reviewer (C.P.) and checked by a second (D. McM). The following information was extracted from primary studies: sample characteristics, sample size, prevalence of major depression, GDS characteristics, diagnostic gold-standard reference test used, data regarding sensitivity and specificity to construct a 2×2 contingency table and quality assessment criteria.

Assessment of quality

The quality of included studies was assessed using the QUADAS-2 (Whiting *et al.*, 2011). Each study was assessed on four domains: participant selection, index test, reference test and flow/timing.

Data synthesis and statistical analysis

For each primary study, a 2×2 contingency table was constructed for every cut-off score reported. This categorises participants into true positives, false negatives, true negatives and false positives according to the GDS score result compared with the gold-standard diagnostic reference test.

The statistical computer software programme STATA was used for data analysis. Pooled estimates of sensitivity, specificity, positive likelihood ratio, negative likelihood ratio and diagnostic odds ratio were calculated using bivariate diagnostic meta-analysis.

Between-study heterogeneity was explored using the I^2 statistic for the pooled diagnostic odds ratio for different cut-off scores of versions of the GDS. Heterogeneity was considered low if the I^2 statistic was 25%, moderate

if 50% and high if 75%. Between-study heterogeneity was explored further if the I^2 statistic was $\geq 50\%$; pooled sensitivity, specificity and diagnostic odds ratios were recalculated excluding potential outliers.

Subgroup and sensitivity analyses were pre-specified in the protocol. Subgroup analysis included exploration of the influence of study setting, country of study (Western versus non-Western) and mean study participant age (according to the definitions of young old (65–74 years), middle old (75–84 years) and old old (>85 years)). Sensitivity analysis included prevalence of major depression and use of a longer GDS version with extraction of a briefer version. Finally, a meta-regression analysis of the logit diagnostic odds ratio was performed to identify sources of heterogeneity. The predictive values used were study setting, country of study, language of GDS, self-administration of the GDS, extraction of a briefer GDS from the GDS-30, average participant age and proportion women.

Results

Figure 1 provides a summary of the selection of studies. The search strategy identified 11,418 records,

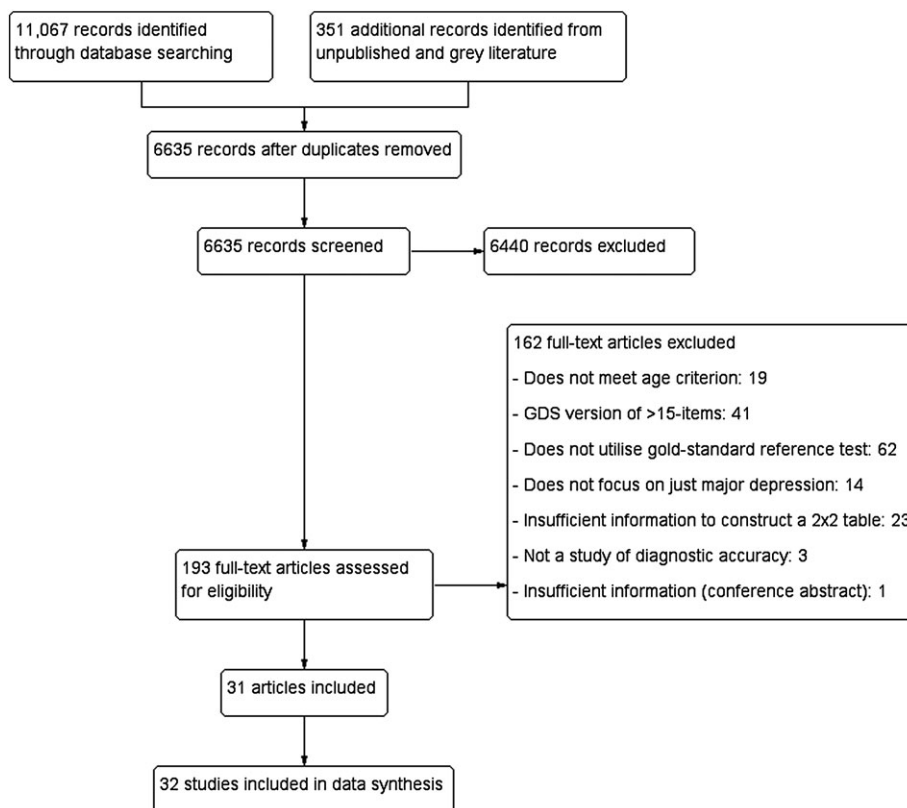


Figure 1 Prisma diagram. GDS, Geriatric Depression Scale.

which resulted in 6635 post-deduplication. One hundred ninety-three records met initial inclusion criteria on the basis of screening titles and abstracts. Full-text copies of these were obtained and examined. Of this 193, 162 studies were then excluded. The remaining 31 records resulted in 32 independent samples; two studies (Allgaier *et al.*, 2011; Broekman *et al.*, 2011) both have two corresponding papers (Nyunt *et al.*, 2009; Allgaier *et al.*, 2013) that together provide complete information for the study in question. Two papers (Blank *et al.*, 2004; Wongpakaran *et al.*, 2013) provide separate sensitivity and specificity data for different settings, and therefore, each setting has been treated as a separate sample. The 32 samples amount to 13,141 participants. None of the samples were from unpublished or grey literature.

Ten of the studies were based in a primary care setting, seven in secondary care, 14 in the community (eight of which were in either a nursing or residential home) and one in a mixed setting (i.e. a combination of community and day hospital) (Table 1). The mean age of the samples ranged from 66.4 to 87.0 years. Twelve of the studies did not assess cognitive functioning, nine studies excluded anyone who had cognitive impairment, six studies specified inclusion criteria as scoring above a certain cut-off score on a cognitive test (i.e. >15 on the mini-mental state examination (MMSE) (three studies), >18 on the MMSE (two studies) and >6 on the abbreviated mental test score (one study) and five studies simply measured cognitive functioning (Table 1)). A recent meta-analysis established the point prevalence of major depression in older adults to be 7.2% (Luppa *et al.*, 2012). The prevalence of major depression in the included studies ranged from 3.2% to 64.1%; nine studies reported the prevalence of major depression to be $\leq 10\%$; 15 studies reported a prevalence of >10–20%, and seven studies reported a prevalence >20% (D'Ath *et al.*, 1994; Gerety *et al.*, 1994; Almeida and Almeida, 1999; Bae and Cho, 2004; Bijl *et al.*, 2006). The studies by Almeida and Bae, which were based on a mental health outpatient clinics, had the highest prevalence of major depression.

The DSM or ICD-10 diagnoses of major depression were included in all 32 studies. Gold-standard diagnostic tests used in the studies included the SCID, MINI, GMS AGECAT, DIS, SCAN, GMS, CIDI, PRIME-MD and ICD-1-checklist.

The GDS-15 was the most common version of the GDS used; all but one study provided sensitivity and specificity data regarding the GDS-15. Nine of these 29 studies extracted data regarding the GDS-15 from the GDS-30. Briefer versions of the GDS used included the GDS-1 (four studies), GDS-4 (five studies), GDS-5

(one study), GDS-7 (one study), GDS-8 (one study) and GDS-10 (five studies). All shorter versions were embedded and extracted from either the GDS-15 or GDS-30.

There is no set standard item(s) for briefer versions of the GDS. See Appendix 2 for a description of items used for briefer versions of the GDS.

In 21 studies, the GDS was orally administered to participants. In nine studies, the GDS was self-administered; in four studies, assistance was available if required. Administration mode was unclear in two studies (McCabe *et al.*, 2006 and Julian *et al.*, 2009). The majority of studies (17) used the English version of the GDS; whereas the remaining studies used translated versions. Of the translated versions, Dutch was the most common language. Remaining languages included Spanish, Portuguese, Korean, Thai, German and Farsi. One study used a mixture of different languages.

Quality assessment

See Table 2 and Appendix 3 for QUADAS-II results for each primary study. The overall rating of risk of bias concerning the GDS was varied; interpretation of the GDS was not blind to gold-standard diagnostic test in three studies, and blinding was unclear in a further six. Not all studies used pre-specified or multiple cut-off scores for the GDS. There was less variation between studies for the overall rating of risk of bias concerning the gold-standard diagnostic reference test. It was unclear in six studies as to whether the gold-standard reference test had been interpreted blind to the results of the GDS. The majority of studies presented a low level of overall risk in terms of flow/timing of study methodology. However, in five studies, not all the participants received the gold-standard diagnostic test; the gold-standard reference test was administered to participants scoring positive on the GDS and only a proportion of those who scored negative in these studies.

Narrative analysis

It was possible to perform a meta-analysis for the GDS-15 but not for other brief versions of the GDS. At least four studies are required to conduct a diagnostic meta-analysis using STATA. For the GDS-5, GDS-7 or GDS-8, there were too few studies. Although there were four or more studies each for the GDS-1, GDS-4 and GDS-10, the items contributing to each of these briefer versions differed within a particular measure

Table 1 All included studies

Study	Sample characteristics (country, setting, age and sex)	Sample size and % depressed	GDS characteristics	Diagnostic standard
Abas <i>et al.</i> (1998)	Country: UK Setting: primary care Age (years): Av. = 68.3 Ethnicity: African-Caribbean Female: 54.0% Cognition: 45% impaired (2% MMSE \leq 9)	N = 164 Major depression: 20.0%	Version: 15 Administration mode: oral Administered by: interviewer Language: English	ICD GMS AGE-CAT
Allgaier <i>et al.</i> (2013)	Country: Germany Setting: community, nursing home Age (years): Av. = 84.5 (range 65–97) Ethnicity: not described Female: 73.9% Cognition: MMSE: \geq 15 for inclusion. Mean MMSE 24.0	N = 92 Major depression: 14.1%	Versions: 15 with 8 and 4 subsets Administration mode: self-administration with assistance if required Administered by: not stated Language: German	DSM-IV SCID
Almeida <i>et al.</i> (1999)	Country: Brazil Setting: secondary care, mental health outpatient clinic Age (years): Av. = 67.5 Ethnicity: not described Female: 84.4% Cognition: Mean MMSE 25.3	N = 64 Major depression: 64.1%	Version: 15 with 10, 4 and 1 subsets Administration mode: oral Administered by: research team Language: Portuguese	ICD-10 ICD-10 checklist of symptoms
Arthur <i>et al.</i> (1999)	Country: UK Setting: primary care Age (years): Av. = 80.0 (range 77–83) Ethnicity: not described Female: 59% Cognition: median CAPE IO score 10	N = 201 Major depression: 6.0%	Version: 15 Administration mode: oral Administered by: practice nurse Language: English	ICD-10 SCAN
Bae <i>et al.</i> (2004)	Country: South Korea Setting: secondary care, mental health outpatient clinic Age (years): Av. = 69.6 Ethnicity: not described	N = 154 Major depression: 40.1%	Versions: 30 with 15 subset Administration mode: self-administration with assistance if required Administered by: research assistance if required Language: Korean	DSM-III-R DIS

(Continues)

Table 1. (Continued)

Study	Sample characteristics (country, setting, age and sex)	Sample size and % depressed	GDS characteristics	Diagnostic standard
Bijl <i>et al.</i> (2006)	Female: 65.0% MMSE ≥ 15 for inclusion Country: Netherlands Setting: primary care Age (years): Av. = 66.5 Ethnicity: not described Female: 64.2% Cognition: MMSE > 18 for inclusion Country: USA Setting: community, nursing home Age (years): Av. = 77.0 Ethnicity: 100% White Female: 67.0% Cognition: cognitive impairment excluded	N = 312 Major depression: 37.5%	Version: 15 Administration mode: oral Administered by: research assistant Language: Dutch Version: 30 with 15 subset Administration mode: oral Administered by: research team Language: English	DSM-IV PRIME-MD
Blank <i>et al.</i> (2004)	Country: USA Setting: community, nursing home Age (years): Av. = 77.0 Ethnicity: 100% White Female: 67.0% Cognition: cognitive impairment excluded	N = 85 Major depression: 9.0%	Version: 30 with 15 subset Administration mode: oral Administered by: research team Language: English	DSM-IV DIS
Blank <i>et al.</i> (2004)	Country: USA Setting: secondary care, inpatients Age (years): Av. = 80.0 Ethnicity: 93.0% White Female: 51.0% Cognition: cognitive impairment excluded	N = 150 Major depression: 8.0%	Version: 30 with 15 subset Administration mode: oral Administered by: research team Language: English	DSM-IV DIS
Blank <i>et al.</i> (2004)	Country: USA Setting: secondary care, outpatient clinic Age (years): Av. = 76.8 Ethnicity: 90.0% White Female: 76.0% Cognition: cognitive impairment excluded	N = 125 Major depression: 11.0%	Version: 30 with 15 subset Administration mode: oral Administered by: research team Language: English	DSM-IV DIS
Broekman <i>et al.</i> (2011)	Country: Singapore Setting: community, social service users Age (years): Av. = 73.8 Ethnicity: 90.1% Chinese, 9.9% Malays and Indians Female: 59.0% Cognition: cognitive impairment excluded	N = 4253 Major depression: 3.4%	Versions: 15 with 7 subset Administration mode: oral Administered by: nurses Languages: English, Chinese and Malay	DSM-IV SCID

(Continues)

Table 1. (Continued)

Study	Sample characteristics (country, setting, age and sex)	Sample size and % depressed	GDS characteristics	Diagnostic standard
Castello <i>et al.</i> (2010)	Country: Brazil Setting: primary care Age (years): 59.5% 60–69 and 40.5% 70–79 Ethnicity: not described Female: 72.7% Cognition: not assessed	N = 220 Major depression: 14.0%	Versions: 30 with 15, 10, 4 and 1 subsets Administration mode: oral Administered by: medical students Language: Spanish	DSM-IV SCID
Cullum <i>et al.</i> (2006)	Country: UK Setting: secondary care, inpatients Age (years): Av. = 80.2 Ethnicity: not described Female: 59% Cognition: AMTS ≥6 for inclusion	N = 221 Major depression: 17.7%	Version: 15 Administration mode: oral Administered by: doctor Language: English	ICD-10 GMS
D'Ath <i>et al.</i> (1994)	Country: UK Setting: primary care Age (years): Av. = 74.4 (range 65–92) Ethnicity: not described Female: 68.3% Cognition: not assessed	N = 120 Major depression: 34.0%	Versions: 15 with 10, 4 and 1 subsets Administration mode: oral Administered by: doctor Language: English	ICD-10 GMS
Davison <i>et al.</i> (2009)	Country: Australia Setting: community, residential home Age (years): Av. = 84.7 (range 67–97) Ethnicity: not described Female: 76.8% Cognition: cognitive impairment excluded	N = 168 Major depression: 16.1%	Version: 15 Administration mode: oral Administered by: research assistant Language: English	DSM-IV SCID
de Craen <i>et al.</i> (2003)	Country: the Netherlands Setting: community Age (years): Av. = 87.0 (range 86–88) Ethnicity: not described Female: 70.0% Cognition: 20% MMSE 0–18, 42% 19–27, 35% 28–30, 3% unknown	N = 79 Major depression: 10.0%	Version: 15 Administration mode: oral Administered by: interviewer Language: Dutch	ICD GMS AGECAT
Friedman <i>et al.</i> (2005)	Country: USA Setting: primary care	N = 960 Major depression: 12.9%	Version: 15 Administration mode: oral	DSM-IV MINI

(Continues)

Table 1. (Continued)

Study	Sample characteristics (country, setting, age and sex)	Sample size and % depressed	GDS characteristics	Diagnostic standard
Gerety <i>et al.</i> (1994)	Age (years): Av. = 79.3 Ethnicity: 97% White Female: 58.2% Cognition: cognitive impairment excluded Country: USA Setting: community, nursing home Age (years): Av. = 78.9 Ethnicity: 74% White Female: 56.0% Cognition: MMSE > 15 for inclusion Country: Spain	N = 134 Major depression: 26.0%	Administered by: interviewer Language: English Versions: 30 with 15 subset Administration mode: oral Administered by: research assistant Language: English	DSM-IV SCID
Izal <i>et al.</i> (2010)	Setting: mixed (community and day hospital) Age (years): Av. = 74.5 Ethnicity: not described Female: 69.0% Cognition: cognitive impairment excluded Country: USA Setting: community COPD patients Age (years): Av. = 66.4	N = 233 Major depression: 11.6%	Versions: 30 with 15, 10 and 5 subsets Administration mode: oral Administered by: psychologist Language: Spanish	DSM-IV SCID
Julian <i>et al.</i> (2009)	Ethnicity: 91.5% white Female: 60.1% Cognition: not assessed Country: Korea Setting: community Age (years): Av. = 72.1	N = 188 Major depression: 11.2%	Version: 15 Administration mode: unclear Administered by: unclear Language: English	DSM-IV MINI
Lee <i>et al.</i> (2013)	Ethnicity: not described Female: 58.3% Cognition: not assessed Country: the Netherlands Setting: primary care Age (years): 43.2% 55–64, 30.7% 65–74, 26.1% ≥75	N = 1941 Major depression: 3.2%	Version: 15 Administration mode: oral Administered by: nurses, social workers and medical students Language: Korean	ICD-10 K-CIDI
Licht-Strunk <i>et al.</i> (2005)		N = 948 Major depression: 13.7%	Version: 15 Administration mode: self-administration Administered by: n/a	DSM RIME-MD

(Continues)

Table 1. (Continued)

Study	Sample characteristics (country, setting, age and sex)	Sample size and % depressed	GDS characteristics	Diagnostic standard
Lyness <i>et al.</i> (1997)	Ethnicity: not described Female: 64.5% Cognition: not assessed Country: USA Setting: primary care	N = 130 Major depression: 9.2%	Language: Dutch Version: 30 with 15 subset Administration mode: self-administration with assistance if required Administered by: n/a Language: English	DSM-III SCID
Malakouti <i>et al.</i> (2006)	Age (years): Av. = 71.0 Ethnicity: 97.7% White, 2.3% Black Female: 58.5% Cognition: not assessed Country: Iran Setting: community	N = 204 Major depression: 10.7%	Version: 15 Administration mode: oral Administered by: psychologist and psychiatrist Language: Farsi	ICD-10 CIDI
Marc <i>et al.</i> (2008)	Ethnicity: not described Female: 53.4% Cognition: not assessed Country: USA Setting: community, nursing home Age (years): Av. = 78.3	N = 526 Major depression: 15.4%	Version: 15 Administration mode: oral Administered by: research assistant Language: English	DSM-IV SCID
McCabe <i>et al.</i> (2006)	Ethnicity: White 85%, Black 11%, Hispanic 4% Female: 65.1% Cognition: MMSE ≥ 18 for inclusion Country: Australia Setting: community, cognitively impaired nursing home residents Age (years): Av. = 86.6 (range 65–99), 89.4% ≥ 80	N = 113 Major depression: 17.7%	Version: 15 Administration mode: unclear Administered by: research assistant Language: English	DSM-IV SCID
Neal and Baldwin (1994)	Ethnicity: not described Female: 74.0% Cognition: 11.5% mildly impaired, 25.0% moderately impaired Country: UK	N = 45	Versions: 30 and 15 subset	ICD

(Continues)

Table 1. (Continued)

Study	Sample characteristics (country, setting, age and sex)	Sample size and % depressed	GDS characteristics	Diagnostic standard
Phelan <i>et al.</i> (2010)	Setting: secondary care, outpatient clinic Age (years): Av. = 77.2 (range 65–90) Ethnicity: not described Female: 62.0% Cognition: not assessed Country: USA Setting: primary care	Major depression: 17.8% N = 69 Major depression: 11.5%	Administration mode: self-administered Administered by: n/a Language: English Version: 15 Administration mode: self-administration with assistance if required Administered by: research assistant if required Language: English	GMS AGECAT DSM-IV SCID
Rait <i>et al.</i> (1999)	Age (years): Av. = 78.0 Ethnicity: 32% non-White Female: 62.0% Cognition: not assessed Country: UK Setting: community	N = 130 Major depression: 10.0%	Version: 15 Administration mode: oral Administered by: research interviewers Language: English	ICD GMS AGECAT
Van Marwijk <i>et al.</i> (1995)	Age (years): Av. = 69.1 Ethnicity: African-Caribbean Female: 50% Cognition: not assessed Country: the Netherlands Setting: primary care	N = 586 Major depression: 5.6%	Versions: 30 with 15, 10, 4 and 1 subsets Administration mode: self-administered Administered by: n/a Language: Dutch	DSM-IV DIS
Watson <i>et al.</i> (2004)	Age (yrs): 59.9% 65–74, 40.1% 75–94 Ethnicity: not described Female: 59.5% Cognition: not assessed Country: USA Setting: community, residential home Age (years): Av. = 83.0 (range 65–100) Ethnicity: not described Female: 72.0% Cognition: not assessed Country: Thailand	N = 112 Major depression: 14.0%	Version: 15 Administration mode: oral Administered by: unclear?	DSM-IV SCID
Wongpakaran <i>et al.</i> (2013)	Setting: secondary care, outpatient clinic Age (years): Av. = 68.8	N = 156 Major depression: 43.6%	Language: English Version: 15 Administration mode: self-administered Administered by: n/a	DSM-IV MINI

(Continues)

Table 1. (Continued)

Study	Sample characteristics (country, setting, age and sex)	Sample size and % depressed	GDS characteristics	Diagnostic standard
Wongpakaran <i>et al.</i> (2013)	Ethnicity: not described Female: 67.9% Cognition: not assessed Country: Thailand Setting: community, nursing home Age (years): Av. = 76.5 Ethnicity: not described Female: 55.6% Cognition: not assessed	N = 81 Major depression: 28.4%	Language: Thai Version: 15 Administration mode: self-administrated Administered by: n/a Language: Thai	DSM-IV MINI

MMSE, mini-mental state examination; CAPE IO: Clifton assessment procedures for the elderly information/orientation; AMTS, abbreviated mental test score.

(Table 3). It was not possible, therefore, to conduct a diagnostic meta-analysis because while a measure in one study may have shared the same name with that used in another, they were essentially different measures.

GDS-1. Four studies reported diagnostic data concerning the GDS-1 (Table 3). The three studies using the same item reported varying sensitivities ranging from 0.18 to 0.62 (Almeida and Almeida, 1999; Castello *et al.*, 2010; Van Marwijk *et al.*, 1995). Specificities reported were similar; ranging from 0.91 to 0.96. D'Ath *et al.* (1994), who used a different item for their GDS-1, reported a sensitivity within the same range (0.59) but a much worse specificity (0.75).

GDS-4. Five studies reported diagnostic data for the GDS-4 (Table 3). Diagnostic data were available for cut-off scores of 1 and 2. The GDS items comprising the GDS-4 for the studies by Allgaier *et al.* (2011, 2013) and D'Ath *et al.* (1994) were the same. At a cut-off score of 1, reported sensitivities ranged from 0.85 to 0.93. Reported specificities ranged from 0.53 to 0.63. The studies by Castello *et al.* (2010) and Van Marwijk *et al.* (1995) used different items for the GDS-4. At a cut-off score of 1, reported sensitivities were lower and ranged from 0.61 to 0.84; however, 95% confidence intervals (CIs) overlapped with the other studies suggesting differences are not significant. Reported specificities were higher and ranged from 0.72 to 0.75, again 95% CIs overlapped with the other studies.

At a cut-off score of 2, the studies by Allgaier *et al.* and D'Ath *et al.* reported sensitivities ranging from 0.54 to 0.61. Reported specificities ranged from 0.89 to 0.92. For the studies by Almeida *et al.*, Castello *et al.* and Van Marwijk *et al.*, reported sensitivities ranged from 0.54 to 0.81 and specificities ranged from 0.66 to 0.94. The 95% CIs for study sensitivities and specificities all overlapped (Table 3).

GDS-7. One study that used a cut-off score of 2 was found for the GDS-7 (Nyunt *et al.*, 2009; Broekman *et al.*, 2011).

GDS-8. One study that used a cut-off score of 5 was found for the GDS-8 (Allgaier *et al.*, 2011, 2013).

GDS-10. Five studies reported diagnostic data for the GDS-10. The items comprising the GDS-10 varied (Table 3 and Appendix 2). Only one study reported diagnostic data at a cut-off score of 2 (Van Marwijk *et al.*, 1995). At a cut-off score of 3, D'Ath *et al.* and

Table 2 QUADAS-II

Study	Patient selection:				Index test:	
	Consecutive or random sample	Avoid case-control/avoid artificially inflated base	Avoided inappropriate exclusions	Overall risk of bias	GDS interpreted blind to reference test	Threshold pre-specified or multiple cut-offs reported
Abas <i>et al.</i> (1998)	✓	✓	✓	Low	✗	✓
Allgaier <i>et al.</i> (2013)	✓	✓	✓	Low	✓	✓
Almeida and Almeida (1999)	?	✗	✓	High	?	✓
Arthur <i>et al.</i> (1999)	✓	✓	✓	Low	✗	✓
Bae <i>et al.</i> (2004)	✓	✗	✓	High	✓	✓
Bijl <i>et al.</i> (2006)	✓	✓	✓	Low	✓	?
Blank <i>et al.</i> (2004)	✓	✓	✓	Low	✓	✓
Broekman <i>et al.</i> (2011)	✓	✓	✓	Low	✓	✓
Castello <i>et al.</i> (2010)	✓	✓	✗	Unclear	✓	✗
Cullum <i>et al.</i> (2006)	✓	✓	✓	Low	?	✓
D'Ath <i>et al.</i> (1994)	✓	✓	✓	Low	?	✓
Davison <i>et al.</i> (2009)	✓	✓	✓	Low	✓	✓
de Craen <i>et al.</i> (2003)	✓	✓	✓	Low	✓	✓
Friedman <i>et al.</i> (2005)	✓	✓	✓	Low	?	✗
Gerety <i>et al.</i> (1994)	✓	✓	✓	Low	✓	✗
Izal <i>et al.</i> (2010)	✓	✓	✓	Low	✓	✓
Julian <i>et al.</i> (2009)	✓	✓	✓	Low	✓	✓
Lee <i>et al.</i> (2013)	✓	?	✓	Unclear	?	✓
Licht-Strunk <i>et al.</i> (2005)	✓	✓	✓	Low	?	✓
Lyness <i>et al.</i> (1997)	✓	✓	✓	Low	✗	✗
Malakouti <i>et al.</i> (2006)	✓	✓	✓	Low	✓	✓
Marc <i>et al.</i> (2008)	✓	✓	✓	Low	✓	✓
McCabe <i>et al.</i> (2006)	✓	✗	✓	High	✓	✓
Neal and Baldwin (1994)	✓	✓	✓	Low	✓	✓
Phelan <i>et al.</i> (2010)	✓	✓	✓	Low	✓	✓
Rait <i>et al.</i> (1999)	✓	✓	✓	Low	✓	✗
Van Marwijk <i>et al.</i> (1995)	✓	✓	✓	Low	✓	?
Watson <i>et al.</i> (2004)	✓	✓	✓	Low	✓	?
Wongpakaran <i>et al.</i> (2013)	✓	✓	✓	Low	✓	✓

Izal *et al.* reported sensitivities ranging from 0.93 to 1.00 and specificities ranging from 0.63 to 0.82. Almeida *et al.*, Castello *et al.* and Van Marwijk *et al.* reported a lower range of sensitivities, 0.52 to 0.92. The reported range of specificities for Almeida *et al.*, Castello *et al.* and Van Marwijk *et al.* (0.65 to 0.83) was similar to that of D'Ath *et al.* and Izal *et al.* Not all 95% CIs for reported sensitivities and specificities overlapped. Almeida *et al.* and Castello *et al.* reported diagnostic data at a cut-off score of 4; the 95% CIs overlapped.

Meta-analysis

GDS-15. Items comprising the GDS-15 are standardised, and therefore, meta-analysis was possible. Twenty studies, out of a total of 32, using the GDS-15 reported multiple cut-off scores with the remaining 12 reporting only a single cut-off score. Not all studies reported the same cut-off scores. See Table 4 for pooled diagnostic properties of the GDS-15 at

different cut-off scores. The recommended cut-off score for the GDS-15 is 5 (Yesavage and Sheikh, 1986); 23 studies ($n=11,468$ participants) reported diagnostic data for this cut-off score. The pooled sensitivity was 0.89 (95% CI 0.80–0.94), and the pooled specificity was 0.77 (95% CI 0.65–0.86) (Table 4).

Between-study heterogeneity measured by the I^2 statistic was 76.7%. The analysis was re-run excluding three studies (de Craen *et al.*, 2003; Watson *et al.*, 2004; Broekman *et al.*, 2011), all of which had diagnostic odds ratios outside the 95% CI of the pooled diagnostic odds ratio. This resulted in a pooled sensitivity and specificity of 0.89 (95% CI 0.80–0.94) and 0.75 (95% CI 0.61–0.86), respectively. The I^2 statistic fell from 76.7% to 33.2%.

Pooled diagnostic data are available for other cut-off scores of the GDS-15 (Table 4). Compared with a cut-off score of 5, a cut-off score of 4 results in a higher sensitivity and lower specificity: 0.88 (95% CI 0.67–0.96) and 0.86 (95% CI 0.68–0.94), respectively. At a cut-off score of 4, the diagnostic odds ratio of the GDS-15 was 42.05 (95% CI 17.42–101.49), which is

Table 2 (Continued)

Study	Index test:			Reference test:		
	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias	Reference test correctly classifies target condition	Reference test interpreted blind to GDS	If translated, appropriate translation
Abas <i>et al.</i> (1998)	n/a	n/a	Low	✓	✓	n/a
Allgaier <i>et al.</i> (2013)	?	?	Low	✓	✓	?
Almeida and Almeida (1999)	✓	✓	Unclear	✓	?	✓
Arthur <i>et al.</i> (1999)	n/a	n/a	High	✓	?	n/a
Bae <i>et al.</i> (2004)	✓	✓	Low	✓	✓	n/a
Bijl <i>et al.</i> (2006)	✓	✓	Unclear	✓	✓	✓
Blank <i>et al.</i> (2004)	n/a	n/a	Low	✓	✓	n/a
Broekman <i>et al.</i> (2011)	✓	✓	Low	✓	✓	✓
Castello <i>et al.</i> (2010)	✓	x	High	✓	✓	✓
Cullum <i>et al.</i> (2006)	n/a	n/a	Unclear	✓	?	n/a
D'Ath <i>et al.</i> (1994)	n/a	n/a	Unclear	✓	?	n/a
Davison <i>et al.</i> (2009)	n/a	n/a	Low	✓	✓	n/a
de Craen <i>et al.</i> (2003)	?	?	Unclear	✓	✓	?
Friedman <i>et al.</i> (2005)	n/a	n/a	High	✓	?	n/a
Gerety <i>et al.</i> (1994)	n/a	n/a	High	✓	✓	n/a
Izal <i>et al.</i> (2010)	✓	✓	Low	✓	✓	✓
Julian <i>et al.</i> (2009)	n/a	n/a	Low	✓	✓	n/a
Lee <i>et al.</i> (2013)	✓	✓	Unclear	✓	?	✓
Licht-Strunk <i>et al.</i> (2005)	✓	✓	Unclear	✓	?	?
Lyness <i>et al.</i> (1997)	n/a	n/a	High	✓	✓	n/a
Malakouti <i>et al.</i> (2006)	✓	✓	Low	✓	✓	✓
Marc <i>et al.</i> (2008)	n/a	n/a	Low	✓	✓	n/a
McCabe <i>et al.</i> (2006)	n/a	n/a	Low	✓	✓	n/a
Neal and Baldwin (1994)	n/a	n/a	Unclear	✓	✓	n/a
Phelan <i>et al.</i> (2010)	n/a	n/a	Low	✓	✓	n/a
Rait <i>et al.</i> (1999)	n/a	n/a	High	✓	✓	n/a
Van Marwijk <i>et al.</i> (1995)	✓	✓	Unclear	✓	✓	✓
Watson <i>et al.</i> (2004)	n/a	n/a	Unclear	✓	✓	n/a
Wongpakaran <i>et al.</i> (2013)	✓	✓	Low	✓	✓	✓

higher than that found for the recommended cut-off score of 5; however, only 10 studies were included in this meta-analysis.

Subgroup analysis of study setting for a cut-off score of 5 found pooled sensitivity and specificity in primary care were 0.92 (95% CI 0.83–0.96) and 0.63 (95% CI 0.42–0.80), respectively (Table 5). Similar pooled diagnostic data were found for secondary care: pooled sensitivity was 0.93 (95% CI 0.88–0.96), and pooled specificity was 0.70 (95% CI 0.53–0.83). In a community setting, pooled sensitivity was the lower at 0.78 (95% CI 0.45–0.94), whereas pooled specificity was 0.90 (95% CI 0.74–0.96), which is higher than pooled specificities in primary and secondary care.

Subgroup analysis of participant age for a cut-off score of 5 revealed that the age group of 'young old' (i.e. 65–74 years of age) and 'middle old' (i.e. 75–84 years of age) had a similar pooled sensitivity. Pooled specificity was lower in the older age group: pooled specificity for the 'young old' was 0.89 (95% CI 0.65–0.97), and pooled specificity for 'middle old'

was 0.73 (95% CI 0.59–0.83) (Appendix 4). It was not possible to make comparisons against studies with a mean participant age of 'very old' (i.e. ≥ 85 years of age) because there were an insufficient number of primary studies for pooling. There were no studies with a mean age of participants that fell between 55 and 64 years of age.

Subgroup analysis of country where study was undertaken revealed that sensitivity values were broadly comparable, although specificity somewhat lower for Western countries. For non-Western countries, at a cut-off score of 5, the pooled sensitivity was 0.90 (95% CI 0.45–0.99), and the pooled specificity was 0.90 (95% CI 0.59–0.98). For Western countries, the pooled sensitivity was 0.88 (95% CI 0.81–0.93), and the pooled specificity was 0.72 (95% CI 0.61–0.81). The diagnostic odds ratio was 79.66 (95% CI 19.52–325.14) for non-Western countries, and 19.09 (95% CI 13.14–27.75) for Western countries (Appendix 5).

Sensitivity analysis explored risk of bias for methodological domains of the primary studies in

Table 2 (Continued)

Study	Reference test:			Flow/timing:		
	If translated, psychometric properties reported	Overall risk of bias	Interval of two weeks or less	All participants receive same reference test	All participants included in analysis?	Overall risk of bias
Abas <i>et al.</i> (1998)	n/a	Low	✓	X	X	High
Allgaier <i>et al.</i> (2013)	✓	Unclear	✓	✓	✓	Low
Almeida and Almeida (1999)	✓	Unclear	X	✓	✓	High
Arthur <i>et al.</i> (1999)	n/a	Unclear	X	✓	✓	High
Bae <i>et al.</i> (2004)	n/a	Low	✓	✓	✓	Low
Bijl <i>et al.</i> (2006)	✓	Low	✓	✓	✓	Low
Blank <i>et al.</i> (2004)	n/a	Low	✓	✓	✓	Low
Broekman <i>et al.</i> (2011)	✓	Low	✓	✓	✓	Low
Castello <i>et al.</i> (2010)	✓	Low	?	✓	✓	Unclear
Cullum <i>et al.</i> (2006)	n/a	Unclear	?	X	X	High
D'Ath <i>et al.</i> (1994)	n/a	Unclear	?	X	X	High
Davison <i>et al.</i> (2009)	n/a	Low	✓	✓	✓	Low
de Craen <i>et al.</i> (2003)	?	Unclear	✓	✓	✓	Low
Friedman <i>et al.</i> (2005)	n/a	Unclear	?	✓	✓	Unclear
Gerety <i>et al.</i> (1994)	n/a	Low	✓	✓	✓	Low
Izal <i>et al.</i> (2010)	✓	Low	?	✓	✓	Low
Julian <i>et al.</i> (2009)	n/a	Low	✓	✓	✓	Low
Lee <i>et al.</i> (2013)	✓	Unclear	?	✓	✓	Unclear
Licht-Strunk <i>et al.</i> (2005)	?	Unclear	?	X	X	High
Lyness <i>et al.</i> (1997)	n/a	Low	✓	✓	✓	Low
Malakouti <i>et al.</i> (2006)	X	High	✓	X	X	High
Marc <i>et al.</i> (2008)	n/a	Low	✓	✓	✓	Low
McCabe <i>et al.</i> (2006)	n/a	Low	✓	✓	✓	Low
Neal and Baldwin (1994)	n/a	Low	✓	✓	✓	Low
Phelan <i>et al.</i> (2010)	n/a	Low	✓	✓	✓	Low
Rait <i>et al.</i> (1999)	n/a	Low	✓	✓	✓	Low
Van Marwijk <i>et al.</i> (1995)	✓	Low	?	✓	✓	Unclear
Watson <i>et al.</i> (2004)	n/a	Low	✓	✓	✓	Low
Wongpakaran <i>et al.</i> (2013)	✓	Low	?	✓	✓	Unclear

accordance with the QUADAS-2, such as patient selection, use and administration of the index, GDS version test, use and administration of a gold-standard, reference test and flow/timing of study design. For each domain, meta-analysis was re-run excluding primary studies that were rated as having a 'high' or 'unclear' risk of bias.

Data for the recommended cut-off score of 5 for the GDS-15 was used. When the meta-analysis was re-run according to patient selection, there was little change to pooled diagnostic data. For example, the pooled sensitivity remained unchanged at 0.89 (95% CI 0.79–0.95), whereas pooled specificity increased slightly to 0.78 (95% CI 0.64–0.87). Again, when meta-analysis was re-run according to risk of bias in the use and administration of the reference test, pooled diagnostic data were similar: pooled sensitivity of studies was 0.90 (95% CI 0.83–0.94), and pooled specificity was 0.78 (95% CI 0.66–0.87). Similar pooled diagnostic data were obtained for meta-analysis of studies rated as 'low' risk of bias for flow/timing of study design; the new pooled sensitivity was 0.90 (95% CI

0.79–1.00), whereas the new pooled specificity was 0.75 (95% CI 0.60–0.85).

When meta-analysis was re-run excluding primary studies rates as having a 'high' or 'unclear' risk of bias, in accordance with the use and administration of the index GDS test, pooled sensitivity increased slightly from 0.89 (95% CI 0.80–0.94) to 0.94 (95% CI 0.86–0.97). Pooled specificity remained relatively unchanged at 0.76 (95% CI 0.60–0.88).

An analysis was also performed to explore the effect of extraction of results for the GDS-15 from an administered GDS-30. At a cut-off score of 5, pooled sensitivity of an extracted GDS-15 and an administered GDS-15 were similar: 0.92 (95% CI 0.85–0.96) and 0.88 (95% CI 0.76–0.94), respectively. Pooled specificity values were also similar: 0.77 (95% CI 0.62–0.88) and 0.78 (95% CI 0.62–0.88), respectively.

Meta-regression. Meta-regression was performed to further explore between-study heterogeneity for the GDS-15 at a cut-off score of 5. Meta-analysis revealed that country (i.e. non-Western) ($p=0.005$) and

Table 3 Pooled diagnostic data for ultra-brief versions of the GDS

Version	Cut-off score	Utilise same items	Study	Sensitivity (95% CI)	Specificity (95% CI)
1	n/a	1	Almeida and Almeida (1999)	0.62 (0.45–0.76)	0.91 (0.72–0.99)
			Castello <i>et al.</i> (2010)	0.48 (0.30–0.67)	0.96 (0.93–0.99)
			Van Marwijk <i>et al.</i> (1995)	0.18 (0.07–0.36)	0.92 (0.90–0.94)
4	1	3	D'Ath <i>et al.</i> (1994)	0.59 (0.42–0.74)	0.75 (0.64–0.84)
			Allgaier <i>et al.</i> (2013)	0.85 (0.55–0.98)	0.53 (0.42–0.65)
			D'Ath <i>et al.</i> (1994)	0.93 (0.80–0.98)	0.63 (0.52–0.74)
		1, 3, 6 and 7	Castello <i>et al.</i> (2010)	0.84 (0.66–0.95)	0.75 (0.68–0.81)
			Van Marwijk <i>et al.</i> (1995)	0.61 (0.42–0.77)	0.72 (0.68–0.76)
			Allgaier <i>et al.</i> (1999)	0.54 (0.25–0.81)	0.92 (0.84–0.97)
	2	1, 3, 6 and 7	D'Ath <i>et al.</i> (1994)	0.61 (0.45–0.76)	0.89 (0.80–0.95)
			Almeida and Almeida (1999)	0.81 (0.65–0.91)	0.78 (0.56–0.93)
			Castello <i>et al.</i> (2010)	0.54 (0.36–0.73)	0.94 (0.90–0.97)
		1, 2, 7 and 9	Van Marwijk <i>et al.</i> (1995)	0.67 (0.48–0.82)	0.66 (0.62–0.70)
			Izal <i>et al.</i> (2010)	0.67 (0.46–0.84)	0.78 (0.72–0.84)
			Broekman <i>et al.</i> (2011)	0.93 (0.88–0.97)	0.91 (0.90–0.92)
5	2	1, 4, 8, 9, and 12	Almeida and Almeida (1999)	0.92 (0.64–0.99)	0.65 (0.53–0.75)
			Castello <i>et al.</i> (2010)	0.77 (0.59–0.90)	0.81 (0.75–0.86)
7	2	1, 3, 4, 5, 7, 8 and 15	Van Marwijk <i>et al.</i> (1995)	0.52 (0.34–0.69)	0.83 (0.80–0.86)
8	5	1, 3, 4, 5, 7, 8, 11 and 14	Almeida and Almeida (1999)	0.85 (0.55–0.98)	0.79 (0.68–0.87)
10	2	1, 2, 4, 5, 7, 8, 9, 12, 13 and 15	Castello <i>et al.</i> (2010)	0.77 (0.59–0.90)	0.81 (0.75–0.86)
			D'Ath <i>et al.</i> (1994)	0.93 (0.80–0.98)	0.63 (0.52–0.74)
			Izal <i>et al.</i> (2010)	1.00 (0.88–1.00)	0.82 (0.76–0.86)
			Almeida and Almeida (1999)	0.92 (0.64–0.99)	0.65 (0.53–0.75)
	3	1, 2, 3, 6, 7, 8, 10, 13, 14 and 15	Castello <i>et al.</i> (2010)	0.77 (0.59–0.90)	0.81 (0.75–0.86)
			Van Marwijk <i>et al.</i> (1995)	0.52 (0.34–0.69)	0.83 (0.80–0.86)
			Almeida and Almeida (1999)	0.85 (0.55–0.98)	0.79 (0.68–0.87)
			Castello <i>et al.</i> (2010)	0.65 (0.45–0.81)	0.89 (0.84–0.93)

GDS, Geriatric Depression Scale; CI, confidence interval.

language (i.e. non-English) ($p=0.05$) was predictive of diagnostic accuracy. Diagnostic accuracy was not influenced by study setting (i.e. primary care versus non-primary care, $p=0.66$), self-administration of the GDS ($p=0.80$), extraction of the GDS-15 from the GDS-30 ($p=0.95$), average age ($p=0.11$) or proportion women ($p=0.54$).

Discussion

This systematic review aimed to establish the diagnostic accuracy of brief versions of the widely used GDS. Existing systematic reviews have focused mainly on the original 30-item version with less attention paid to briefer versions, particularly those using fewer than 15 items, which may be more suitable for clinical practice.

This systematic review found a sensitivity of 0.89 and a specificity of 0.77 for the GDS-15 at the recommended cut-off score of 5. Comparison of findings with previous reviews is difficult because of how data for different cut-off scores have been pooled. One previous review (Dennis *et al.*, 2012) reports sensitivity and specificity for the GDS-15 at a cut-off score of 5; the sensitivity and specificity established in this review are higher.

At a cut-off score of 4, diagnostic data were more favourable: sensitivity was 0.88, and specificity was

0.86, which resulted in a greater diagnostic odds ratio compared with a cut-off score of 5 (42.05 and 27.28, respectively). At a cut-off score of 6, sensitivity (0.80) was lower than that found for a cut-off score of 5, although specificity (0.83) was higher. After a cut-off score of 5, pooled sensitivity consistently fell and pooled specificity consistently increased. Pooled diagnostic data for cut-off scores other than 5 have to be interpreted cautiously because fewer studies are included in meta-analysis.

Several briefer versions of the GDS were found: GDS-1, GDS-4, GDS-5, GDS-7, GDS-8 and GDS-10. However, there was inconsistency in the items that contributed to these briefer versions, and there are no standardised cut-off scores. Meta-analyses could not be performed for any briefer versions of the GDS because of an inadequate number of studies for the different cut-off scores reported. Therefore, it is difficult to comment on which briefer version of the GDS performs best.

Limitations

Limitations in this review refer to the primary studies included and also refer directly to the review itself. The most important limitation related to the primary studies is the possibility of selective reporting of cut-off

Table 4 Pooled diagnostic data for the GDS-15 at different cut-off scores

Cut-off score	No. of studies	N	Prev. of Maj Dep (%)	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)
2	4	1517	9.8	0.90 (0.79–0.95)	0.43 (0.35–0.51)	1.57 (1.41–1.74)	0.25 (0.12–0.46)	6.38 (3.34–12.20)
3	6	5849	10.7	0.95 (0.77–0.99)	0.68 (0.57–0.77)	2.96 (2.15–4.06)	0.07 (0.01–0.39)	42.04 (6.58–268.52)
4	10	7874	10.1	0.88 (0.67–0.96)	0.86 (0.68–0.94)	6.06 (2.78–13.24)	0.14 (0.05–0.39)	42.05 (17.42–101.49)
5	23	11,468	11.5	0.89 (0.80–0.94)	0.77 (0.65–0.86)	3.93 (2.58–6.00)	0.14 (0.09–0.24)	27.28 (16.57–44.93)
6	20	9886	11.8	0.79 (0.68–0.87)	0.83 (0.72–0.90)	4.53 (2.85–7.20)	0.26 (0.17–0.38)	17.61 (10.12–30.63)
7	11	8678	10.7	0.72 (0.53–0.85)	0.90 (0.80–0.95)	6.91 (3.95–12.10)	0.32 (0.19–0.54)	21.73 (12.67–37.28)
8	9	7541	10.3	0.70 (0.43–0.88)	0.91 (0.78–0.97)	7.84 (3.69–16.67)	0.33 (0.16–0.67)	23.90 (10.84–52.72)
9	8	3321	9.4	0.52 (0.30–0.72)	0.92 (0.83–0.96)	6.36 (4.08–9.91)	0.52 (0.34–0.79)	12.20 (8.09–18.39)
10	6	3127	9.2	0.47 (0.27–0.69)	0.94 (0.87–0.97)	8.10 (5.3–12.4)	0.56 (0.38–0.82)	14.00 (10.00–21.00)

GDS, Geriatric Depression Scale; CI, confidence interval.

Table 5 Pooled diagnostic data for the GDS-15 at different cut-off scores in different settings

Cut-off score	Setting	No. of studies	N	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)
5	Primary	9	3124	0.92 (0.83–0.96)	0.63 (0.42–0.80)	2.49 (1.54–4.03)	0.13 (0.08–0.21)	18.58 (13.14–26.27)
	Secondary Community	5	640	0.93 (0.88–0.96)	0.70 (0.53–0.83)	3.05 (1.85–5.04)	0.11 (0.07–0.17)	29.00 (13.27–63.38)
6	Primary	8	7471	0.78 (0.45–0.94)	0.90 (0.74–0.96)	7.36 (3.22–16.83)	0.25 (0.09–0.72)	29.31 (9.19–93.47)
	Secondary	6	1734	0.77 (0.69–0.84)	0.74 (0.47–0.90)	2.93 (1.26–6.81)	0.31 (0.21–0.46)	9.44 (3.01–29.58)
	Community	5	649	0.86 (0.80–0.90)	0.76 (0.63–0.86)	3.64 (2.25–5.88)	0.18 (0.13–0.26)	19.81 (9.87–39.74)
	Community	9	7503	0.80 (0.54–0.93)	0.89 (0.76–0.96)	7.54 (3.62–15.73)	0.22 (0.09–0.56)	33.83 (14.26–80.23)

GDS, Geriatric Depression Scale; CI, confidence interval.

scores. Selective reporting is suggested by two means. Firstly, 32 studies that used the GDS-15 were identified, but only 23 studies reported diagnostic data at the recommended cut-off score of 5. Secondly, the expected changes in sensitivity and specificity as cut-off score of the GDS-15 rises are not observed. As the cut-off score increases, sensitivity should fall, and specificity should rise. Table 4 illustrates that pooled sensitivity rises at a cut-off score of 3 and then again at a cut-off score of 5. Pooled specificity rises to a cut-off score of 4 then drops at a cut-off score of 5. Pooled specificity at a cut-off score of 6 is lower than that found at a cut-off score of 4. One interpretation of these findings is that the decision to report a cut-off point is determined by its performance in a particular study, with studies more likely to report a cut-off point when it performs well. This places considerable limitations on the results of any diagnostic meta-analysis of the GDS-15 because diagnostic performance is artificially inflated.

A further consideration is that not all studies measured cognitive functioning and some applied exclusion criteria regarding it. The presence of cognitive impairment may substantially affect the diagnostic accuracy of a depression measure in an older adult population. Future studies of the GDS may want to report its diagnostic performance separately for samples including and then excluding people with cognitive deficits.

The primary studies had a number of methodological issues as shown by the results of the QUADAS-II. The QUADAS-II domain of 'index test' (i.e. use and administration of the GDS—interpretation blind to the reference test, pre-specified or multiple cut-offs reported, appropriate translation if applicable and if translated, psychometric properties reported) was of concern. Overall risk of bias of the 'index test' did influence diagnostic performance; when primary studies that were rated as having an overall 'high' or 'unclear' risk of bias were removed from meta-analysis, diagnostic performance improved, which is contrary to what would be expected; it is unclear why this was the case. Bias concerning the QUADAS-II domain of 'patient selection' did not influence pooled diagnostic data; pooled sensitivity and specificity of only primary studies rated as having a 'low' overall risk of bias led to no change in reported sensitivity or specificity. For the QUADAS-II domain of 'flow/timing', an interval of more than 2 weeks between administration of the GDS and reference test did influence pooled diagnostic data; specificity increased (from 0.75 to 0.77) and sensitivity fell (from 0.91 to 0.89) when meta-analysis was re-run excluding primary studies where risk was rated as 'high' or 'unclear'.

The protocol developed was not registered, which is a potential limitation of this review. Although attempts were made to reduce publication bias by searching the grey literature, the abstracting of diagnostic validation studies is variable, so it remains possible that relevant studies were missed. In addition, bias could be introduced by study selection and data extraction being performed by one reviewer despite predefined inclusion and exclusion criteria being followed and applied.

Research implications

This review has highlighted several issues. First, as discussed, our results suggest that there may be selective reporting of results, which limits the interpretation of diagnostic meta-analyses of this measure. We suggest that diagnostic validation studies of the GDS report all cut-off scores to ensure that the selective reporting of cut-off points does not artificially inflate the observed diagnostic accuracy of the GDS. The effect of cognitive impairment on diagnostic accuracy also needs to be addressed in future research. Knowledge concerning the diagnostic accuracy of versions of the GDS with fewer than 15 items is currently limited. Future studies of these briefer versions should also report a range of cut-off points. A range of items are used to contribute to these brief versions. The diagnostic performance of the differently constructed measures should be compared to identify which combination has greatest accuracy.

Conclusions

This review provides information regarding the diagnostic performance of the GDS-15. It is difficult to make firm conclusions because our pooled results show evidence of selective reporting of cut-off scores; therefore, our findings should be interpreted cautiously. Selective reporting of cut-off scores leads to the diagnostic accuracy of the screening instruments being exaggerated because results for cut-off scores that perform less well are not reported.

Briefer versions of the GDS may have more clinical appeal because of the time restraints faced in clinical practice, but unfortunately, meta-analyses were not possible for briefer versions because of an inadequate number of primary studies. The sensitivity and specificity of briefer versions of the GDS need to be explored further so that recommendations can be made.

Conflict of interest

None declared.

Key points

- This review examines the utility of brief versions of the GDS in screening for depression.
- The possibility of selective reporting of cut-off scores means results of meta-analysis should be approached cautiously.
- There is a need for more research using brief versions of the GDS.

References

- Abas MA, Phillips C, Carter J, *et al.* 1998. Culturally sensitive validation of screening questionnaires for depression in older African-Caribbean people living in south London. *Br. J. Psychiatry* **173**: 249–254.
- Allgaier AK, Kramer D, Mergl R, Fejtikova S, Hegerl U. 2011. Validity of the Geriatric Depression Scale in nursing home residents: Comparison of GDS-15, GDS-8, and GDS-4. [German] Validität der Geriatrischen Depressionskala bei Altenheimbewohnern: Vergleich von GDS-15, GDS-8 und GDS-4. *Psychiatr. Prax.* **38**(6): 280–286.
- Allgaier AK, Kramer D, Saravo B, *et al.* 2013. Beside the Geriatric Depression Scale: the WHO-Five Well-being Index as a valid screening tool for depression in nursing homes. *Int. J. Geriatr. Psychiatry* **28**(11): 1197–1204.
- Almeida OP, Almeida SA. 1999. Short versions of the Geriatric Depression Scale: a study of their validity for the diagnosis of a major depressive episode according to ICD-10 and DSM-IV. *Int. J. Geriatr. Psychiatry* **14**(10): 858–865.
- Anderson DN. 2001. Treating depression in old age: the reasons to be positive. *Age Ageing* **30**(1): 13–17.
- Arthur A, Jagger C, Lindesay J, Graham C, Clarke M. 1999. Using an annual over-75 health check to screen for depression: validation of the short Geriatric Depression Scale (GDS15) within general practice. *Int. J. Geriatr. Psychiatry* **14**(6): 431–439.
- Bae JN, Cho MJ. 2004. Development of the Korean version of the Geriatric Depression Scale and its short form among elderly psychiatric patients. *J. Psychosom. Res.* **57**(3): 297–305.
- Bijl D, van Marwijk HW, Adér HJ, Beekman AT, de Haan M. 2006. Test-characteristics of the GDS-15 in screening for major depression in elderly patients in general practice. *Clin. Gerontol.* **29**(1): 1–9.
- Birrer RB, Vemuri SP. 2004. Depression in later life: a diagnostic and therapeutic challenge. *Am. Fam. Physician* **69**(10): 2375–2382.
- Blank K, Gruman C, Robison JT. 2004. Case-finding for depression in elderly people: balancing ease of administration with validity in varied treatment settings. *J. Gerontol. A Biol. Sci. Med. Sci.* **59**(4): 378–384.
- Broekman BFP, Niti M, Nyunt MSZ, *et al.* 2011. Validation of a brief seven-item response bias-free Geriatric Depression Scale. *Am. J. Geriatr. Psychiatry* **19**(6): 589–596.
- Castello MS, Coelho-Filho JM, Carvalho AF, *et al.* 2010. Validity of the Brazilian version of the Geriatric Depression Scale (GDS) among primary care patients. *Int. Psychogeriatr.* **22**(1): 109–113.
- de Craen AJ, Heeren T, Gussekloo J. 2003. Accuracy of the 15-item geriatric depression scale (GDS-15) in a community sample of the oldest old. *Int. J. Geriatr. Psychiatry* **18**(1): 63–66.
- Cullum S, Tucker S, Todd C, Brayne C. 2006. Screening for depression in older medical inpatients. *Int. J. Geriatr. Psychiatry* **21**(5): 469–476.
- D'Ath P, Katona P, Mullan E, Evans S, Cornelius K. 1994. Screening, detection and management of depression in elderly primary-care attenders: the acceptability and performance of the 15-item Geriatric Depression Scale (GDS15) and the development of short versions. *Fam. Pract.* **11**(3): 260–266.
- Davison TE, McCabe MP, Mellor D. 2009. An examination of the gold standard diagnosis of major depression in aged-care settings. *Am. J. Geriatr. Psychiatry* **17**(5): 359–367.
- Dennis M, Kadri A, Coffey J. 2012. Depression in older people in the general hospital: a systematic review of screening instruments. *Age Ageing* **41**(2): 148–154.
- Drayer RA, Mulsant BH, Lenze EJ, *et al.* 2005. Somatic symptoms of depression in elderly patients with medical comorbidities. *Int. J. Geriatr. Psychiatry* **20**(10): 973–82.
- Fiske A, Wetherell JL, Gatz M. 2009. Depression in older adults. *Annu. Rev. Clin. Psychol.* **5**: 363–389.
- Friedman B, Heisel MJ, Delavan RL. 2005. Psychometric properties of the 15-item geriatric depression scale in functionally impaired, cognitively intact, community-dwelling elderly primary care patients. *J. Am. Geriatr. Soc.* **53**(9): 1570–1576.
- Gerety MB, Williams JW Jr, Mulrow CD, *et al.* 1994. Performance of case-finding tools for depression in the nursing home: influence of clinical and functional characteristics and selection of optimal threshold scores. *J. Am. Geriatr. Soc.* **42**(10): 1103–1109.
- Hoyle MT, Alessi CA, Harker JO, *et al.* 1999. Development and testing of a five-item version of the Geriatric Depression Scale. *J. Am. Geriatr. Soc.* **47**: 873–878.
- Izal M, Montorio I, Nuevo R, Perez-Rojo G, Cabrera I. 2010. Optimising the diagnostic performance of the Geriatric Depression Scale. *Psychiatry Res.* **178**(1): 142–146.
- Jongeneel L, Eisses A, Pot A, Beekman A, Ribbe M. 2002. Depression in long term care facilities: validation and reliability of the geriatric depression scale in a Dutch nursing home population. *Gerontologist* **42**: 256–256.
- Julian LJ, Gregorich SE, Earnest G, *et al.* 2009. Screening for depression in chronic obstructive pulmonary disease. *Copd: J. Chron. Obstruct. Pulmon. Dis.* **6**(6): 452–458.
- Lee SC, Kim WH, Chang SM, *et al.* 2013. The use of the Korean version of short form Geriatric Depression Scale (SGDS-K) in the community dwelling elderly in Korea. *J. Korean Geriatr. Psychiatry* **17**(1): 37–43.
- Licht-Strunk E, van der Kooij KG, van Schaik DJF, *et al.* 2005. Prevalence of depression in older patients consulting their general practitioner in the Netherlands. *Int. J. Geriatr. Psychiatry* **20**(11): 1013–1019.
- Luppa M, Sikorski C, Luck T, *et al.* 2012. Age- and gender-specific prevalence of depression in latest-life—systematic review and meta-analysis. *J. Affect. Disord.* **136**(3): 212–21.
- Lyness JM, Noel TK, Cox C, *et al.* 1997. Screening for depression in elderly primary care patients: a comparison of the center for epidemiologic studies—depression scale and the Geriatric Depression Scale. *Arch. Intern. Med.* **157**(4): 449–454.
- Malakouti SK, Fatollahi P, Mirabzadeh A, Salavati M, Zandi T. 2006. Reliability, validity and factor structure of the GDS-15 in Iranian elderly. *Int. J. Geriatr. Psychiatry* **21**(6): 588–593.
- Marc LG, Raue PJ, Bruce ML. 2008. Screening performance of the 15-item Geriatric Depression Scale in a diverse elderly home care population. *Am. J. Geriatr. Psychiatry* **16**(11): 914–921.
- McCabe MP, Davison T, Mellor D, *et al.* 2006. Depression among older people with cognitive impairment: prevalence and detection. *Int. J. Geriatr. Psychiatry* **21**(7): 633–644.
- Mitchell AJ, Bird V, Rizzo M, Meader N. 2010a. Diagnostic validity and added value of the Geriatric Depression Scale for depression in primary care: a meta-analysis of GDS30 and GDS15. *J. Affect. Disord.* **125**(1–3): 10–17.
- Mitchell AJ, Bird V, Rizzo M, Meader N. 2010b. Which version of the Geriatric Depression Scale is most useful in medical settings and nursing homes? Diagnostic validity meta-analysis. *Am. J. Geriatr. Psychiatry* **18**(12): 1066–1077.
- Neal RM, Baldwin RC. 1994. Screening for anxiety and depression in elderly medical outpatients. *Age Ageing* **23**(6): 461–464.
- NICE (2011). Common mental health disorders: identification and pathways to care. NICE clinical guideline 123. Available at www.nice.org.uk/CG123 [NICE guideline].
- NICE (2014). Antenatal and postnatal mental health: clinical management and service guidance. NICE clinical guideline 192. Available at www.nice.org.uk/CG192 [NICE guideline].
- Nyunt MSZ, Fones C, Niti M, Ng TP. 2009. Criterion-based validity and reliability of the Geriatric Depression Screening Scale (GDS-15) in a large validation sample of community-living Asian older adults. *Aging Ment. Health* **13**(3): 376–382.
- Phelan E, Williams B, Meeker K, *et al.* 2010. A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam. Pract.* **11**: 63.
- Rait G, Burns A, Baldwin R, *et al.* 1999. Screening for depression in African-Caribbean elders. *Fam. Pract.* **16**(6): 591–595.
- Rinaldi P, Mecocci P, Benedetti C, *et al.* 2003. Validation of the five-item Geriatric Depression Scale in elderly subjects in three different settings. *J. Am. Geriatr. Soc.* **51**(5): 694–698.
- Van Marwijk HWJ, Wallace P, De Bock GH, *et al.* 1995. Evaluation of the feasibility, reliability and diagnostic value of shortened versions of the Geriatric Depression Scale. *Br. J. Gen. Pract.* **45**(393): 195–199.
- Wancata J, Alexandrowicz R, Marquart B, Weiss M, Friedrich F. 2006. The criterion validity of the Geriatric Depression Scale: a systematic review. *Acta Psychiatr. Scand.* **114**(6): 398–410.
- Watson LC, Pignone MP. 2003. Screening accuracy for late-life depression in primary care: a systematic review. *J. Family Pract.* **52**(12): 956–964.
- Watson LC, Lewis CL, Kistler CE, Amick HR, Boustani M. 2004. Can we trust depression screening instruments in healthy 'old-old' adults? *Int. J. Geriatr. Psychiatry* **19**(3): 278–285.
- Whiting PF, Rutjes AW, Westwood ME, *et al.* 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* **155**(8): 529–536.
- Wongpakaran N, Wongpakaran T, Van Reekum R. 2013. The use of GDS-15 in detecting MDD: a comparison between residents in a Thai long-term care home and geriatric outpatients. *J. Clin. Med. Res.* **5**(2): 101.
- Yesavage JA, Sheikh JL. 1986. Geriatric Depression Scale (GDS) recent evidence and development of a shorter version. *Clin. Gerontol.* **5**(1–2): 165–173.
- Yesavage JA, Brink T, Rose TL, *et al.* 1982. Development and validation of a Geriatric Depression Screening Scale: a preliminary report. *J. Psychiatr. Res.* **17**(1): 37–49.

Appendix 1: MEDLINE search strategy

1. older\$.ti,ab.
2. elder\$.ti,ab.
3. geriatri\$.ti,ab.
4. 1 or 2 or 3
5. Limit 4 to (humans and yr="1982-Current")
6. exp Depression/
7. exp Depressive Disorder/
8. (depressive or depression or depressed).ti,ab.
9. (melancholi\$ or dysphori\$ or dysthymi\$).ti,ab.
10. 6 or 7 or 8 or 9
11. Limit 10 to (humans and yr="1982-Current")
12. "geriatric depression scale".ti,ab.
13. "GDS\$".ti,ab.
14. 12 or 13
15. Limit 14 to (humans and yr="1982-Current")
16. 5 and 11 and 15

Appendix 2: Geriatric Depression Scale (GDS) items

Item number	GDS-15 item	Allgaier <i>et al.</i> (2013)	Almeida and Almeida (1999)	Broekman <i>et al.</i> (2011)	Castello <i>et al.</i> (2010)	D'Ath <i>et al.</i> (1994)	Izal <i>et al.</i> (2010)	Van Marwijk <i>et al.</i> (1995)
1	Are you basically satisfied with your life?	GDS-4 GDS-8	GDS-1 GDS-4 GDS-10	GDS-7	GDS-1 GDS-4 GDS-10	GDS-4 GDS-10	GDS-5 GDS-10	GDS-1 GDS-4 GDS-10
2	Have you dropped many of your activities and interests?		GDS-4 GDS-10		GDS-4 GDS-10	GDS-10	GDS-10	GDS-4 GDS-10
3	Do you feel that your life is empty?	GDS-4 GDS-8		GDS-7		GDS-1 GDS-4 GDS-10	GDS-10	
4	Do you often get bored?	GDS-8	GDS-10	GDS-7	GDS-10		GDS-5	GDS-10
5	Are you in good spirits most of the time?	GDS-8	GDS-10	GDS-7	GDS-10			GDS-10
6	Are you afraid that something bad is going to happen to you?	GDS-4				GDS-4 GDS-10	GDS-10	
7	Do you feel happy most of the time?	GDS-4 GDS-8	GDS-4 GDS-10	GDS-7	GDS-4 GDS-10	GDS-4 GDS-10	GDS-10	GDS-4 GDS-10
8	Do you feel helpless?	GDS-8	GDS-10	GDS-7	GDS-10	GDS-10	GDS-5 GDS-10	GDS-10
9	Do you prefer to stay at home rather than going out and doing new things?		GDS-4 GDS-10		GDS-4		GDS-5	GDS-4 GDS-10
10	Do you feel you have more problems with memory than most?				GDS-10	GDS-10	GDS-10	
11	Do you think it is wonderful to be alive?	GDS-8						
12	Do you feel pretty worthless the way you are now?		GDS-10		GDS-10		GDS-5	GDS-10

(Continues)

Appendix 2: (Continued)

Item number	GDS-15 item	Allgaier <i>et al.</i> (2013)	Almeida and Almeida (1999)	Broekman <i>et al.</i> (2011)	Castello <i>et al.</i> (2010)	D'Ath <i>et al.</i> (1994)	Izal <i>et al.</i> (2010)	Van Marwijk <i>et al.</i> (1995)
13	Do you feel full of energy?		GDS-10		GDS-10	GDS-10	GDS-10	GDS-10
14	Do you feel that your situation is hopeless?	GDS-8				GDS-10	GDS-10	
15	Do you think that most people are better off than you are?		GDS-10	GDS-7	GDS-10	GDS-10	GDS-10	GDS-10

Appendix 3: QUADAS-II

Study	Patient selection: Applicability	Index test: Applicability	Reference test: Applicability
Abas <i>et al.</i> (1998)	x	✓	✓
Allgaier <i>et al.</i> (2013)	✓	✓	✓
Almeida and Almeida (1999)	x	✓	✓
Arthur <i>et al.</i> (1999)	✓	✓	✓
Bae and Cho (2004)	x	✓	✓
Bijl <i>et al.</i> (2006)	✓	✓	✓
Blank <i>et al.</i> (2004)	✓	✓	✓
Broekman <i>et al.</i> (2011)	✓	✓	✓
Castello <i>et al.</i> (2010)	✓	✓	✓
Cullum <i>et al.</i> (2006)	✓	✓	✓
D'Ath <i>et al.</i> (1994)	✓	✓	✓
Davison <i>et al.</i> (2009)	✓	✓	✓
de Craen <i>et al.</i> (2003)	x	✓	✓
Friedman <i>et al.</i> (2005)	✓	✓	✓
Gerety <i>et al.</i> (1994)	✓	✓	✓
Izal <i>et al.</i> (2010)	✓	✓	✓
Julian <i>et al.</i> (2009)	✓	✓	✓
Lee <i>et al.</i> (2013)	✓	✓	✓
Licht-Strunk <i>et al.</i> (2005)	✓	✓	✓
Lyness <i>et al.</i> (1997)	✓	✓	✓
Malakouti <i>et al.</i> (2006)	✓	✓	✓
Marc <i>et al.</i> (2008)	✓	✓	✓
McCabe <i>et al.</i> (2006)	x	✓	✓
Neal and Baldwin (1994)	✓	✓	✓
Phelan <i>et al.</i> (2010)	✓	✓	✓
Rait <i>et al.</i> (1999)	x	✓	✓
Van Marwijk <i>et al.</i> (1995)	✓	✓	✓
Watson <i>et al.</i> (2004)	✓	✓	✓
Wongpakaran <i>et al.</i> (2013)	x	✓	✓

Appendix 4: Pooled diagnostic data for the GDS-15 at different age ranges

Cut-off score	Age	No. of studies	N	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)
5	YO	7	6862	0.87 (0.59-0.97)	0.89 (0.65-0.97)	7.91 (2.46-25.43)	0.15 (0.05-0.48)	53.06 (20.84-135.09)
	MO	10	2655	0.88 (0.76-0.94)	0.73 (0.59-0.83)	3.19 (2.13-4.77)	0.17 (0.09-0.32)	18.72 (9.42-37.02)
6	YO	6	6732	0.74 (0.43-0.92)	0.87 (0.46-0.98)	5.44 (3.86-84.69)	0.30 (0.13-0.70)	18.09 (3.86-84.69)
	MO	10	2538	0.56 (0.30-0.79)	0.55 (0.26-0.81)	1.25 (0.42-3.67)	0.80 (0.27-2.39)	1.56 (0.18-13.76)
7	YO	4	6412	0.67 (0.26-0.92)	0.95 (0.68-0.99)	12.41 (2.76-55.76)	0.35 (0.12-1.01)	35.59 (12.92-98.05)
	MO	5	1841	0.64 (0.50-0.77)	0.88 (0.80-0.93)	5.33 (3.70-7.67)	0.41 (0.29-0.56)	13.16 (9.55-18.15)

GDS, Geriatric Depression Scale; CI, confidence interval.

Appendix 5: Pooled diagnostic data for the GDS-15 in Western and non-Western countries

Cut-off score	Western country	No. of studies	No. of participants	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	Diagnostic odds ratio (95% CI)
4	Yes	6	1320	0.85 (0.75-0.92)	0.77 (0.64-0.87)	0.12 (0.06-0.21)	16.67 (11.20-24.82)	0.01 (0.00-0.02)
	No	4	6554	0.93 (0.33-1.00)	0.93 (0.60-0.99)	3.75 (2.18-6.47)	0.19 (0.10-0.35)	19.75 (6.98-55.82)
5	Yes	18	3130	0.88 (0.81-0.93)	0.72 (0.61-0.81)	12.8 (2.24-73.30)	0.08 (0.00-1.36)	163.51 (47.40-564.00)
	No	5	6708	0.90 (0.45-0.99)	0.90 (0.59-0.98)	3.13 (2.30-4.26)	0.16 (0.11-0.25)	19.09 (13.14-27.75)
6	Yes	15	3178	0.78 (0.71-0.84)	0.78 (0.68-0.86)	9.27 (2.05-42.01)	0.12 (0.02-0.82)	79.66 (19.52-325.14)
	No	5	6708	0.82 (0.40-0.97)	0.93 (0.67-0.99)	3.56 (2.39-5.29)	0.28 (0.21-0.37)	12.57 (7.24-21.82)
7	Yes	7	2126	0.72 (0.60-0.82)	0.84 (0.75-0.90)	4.41 (3.21-6.05)	0.33 (0.24-0.46)	13.20 (10.00-17.42)
	No	4	6552	0.77 (0.27-0.97)	0.94 (0.67-0.99)	13.57 (2.82-65.20)	0.25 (0.05-1.21)	55.38 (16.90-181.42)
8	Yes	4	908	0.54 (0.37-0.70)	0.90 (0.86-0.93)	5.28 (3.95-7.06)	0.52 (0.37-0.72)	10.26 (5.97-17.62)
	No	5	6633	0.80 (0.39-0.96)	0.92 (0.62-0.99)	10.38 (2.23-48.30)	0.22 (0.06-0.81)	46.69 (18.84-115.72)
9	Yes	4	941	0.39 (0.20-0.62)	0.94 (0.91-0.96)	6.09 (3.97-9.34)	0.66 (0.47-0.92)	9.28 (4.50-19.15)
	No	4	2380	0.64 (0.28-0.89)	0.89 (0.63-0.98)	5.91 (2.36-14.77)	0.40 (0.18-0.90)	14.65 (9.12-23.55)

GDS, Geriatric Depression Scale; CI, confidence interval.

BMJ Open Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis

Laura Manea,^{1,2} Jan Rasmus Boehnke,³ Simon Gilbody,^{1,2} Andrew S Moriarty,² Dean McMillan^{1,2}

To cite: Manea L, Boehnke JR, Gilbody S, *et al.* Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis. *BMJ Open* 2017;**7**:e015247. doi:10.1136/bmjopen-2016-015247

► Prepublication history and additional material for this paper are available online. To view, please visit the journal (<http://dx.doi.org/10.1136/bmjopen-2016-015247>).

Received 23 November 2016
Revised 21 July 2017
Accepted 25 July 2017



CrossMark

¹Department of Health Sciences, University of York, York, UK

²Hull York Medical School, University of York, York, United Kingdom

³Dundee Centre for Health and Related Research, University of Dundee, Dundee, United Kingdom

Correspondence to

Dr Laura Manea;
laura.manea@york.ac.uk

ABSTRACT

Objectives To investigate whether an authorship effect is found that leads to better performance in studies conducted by the original developers of the Patient Health Questionnaire (PHQ-9) (allegiant studies).

Design Systematic review with random effects bivariate diagnostic meta-analysis. Search strategies included electronic databases, examination of reference lists and forward citation searches.

Inclusion criteria Included studies provided sufficient data to calculate the diagnostic accuracy of the PHQ-9 against a gold standard diagnosis of major depression using the algorithm or the summed item scoring method at cut-off point 10.

Data extraction Descriptive information, methodological quality criteria and 2×2 contingency tables.

Results Seven allegiant and 20 independent studies reported the diagnostic performance of the PHQ-9 using the algorithm scoring method. Pooled diagnostic OR (DOR) for the allegiant group was 64.40, and 15.05 for non-allegiant studies group. The allegiance status was a significant predictor of DOR variation ($p < 0.0001$). Five allegiant studies and 26 non-allegiant studies reported the performance of the PHQ-9 at recommended cut-off point of 10. Pooled DOR for the allegiant group was 49.31, and 24.96 for the non-allegiant studies. The allegiance status was a significant predictor of DOR variation ($p = 0.015$). Some potential alternative explanations for the observed authorship effect including differences in study characteristics and quality were found, although it is not clear how some of them account for the observed differences.

Conclusions Allegiant studies reported better performance of the PHQ-9. Allegiance status was predictive of variation in the DOR. Based on the observed differences between independent and non-independent studies, we were unable to conclude or exclude that allegiance effects are present in studies examining the diagnostic performance of the PHQ-9. This study highlights the need for future meta-analyses of diagnostic validation studies of psychological measures to evaluate the impact of researcher allegiance in the primary studies.

Research on allegiance effects has a long tradition in psychotherapy research. In this

Strengths and limitations of this study

- An original study—the first meta-analysis of diagnostic validation studies of psychological measures to evaluate the impact of researcher allegiance.
- Using rigorous methodology—strict inclusion/exclusion and quality assessment criteria.
- We found that the allegiance effect was a significant predictor of the variation of the diagnostic OR in the meta-regression analysis.
- Substantial variability observed in methodological quality of included studies.
- Based on the observed methodological differences between the independent and non-independent studies, we were unable to conclude or exclude that allegiance effects are present in studies examining the diagnostic performance of the Patient Health Questionnaire (PHQ-9).

context, *allegiance* describes the phenomenon that researchers and clinicians who developed a treatment approach or are for other reasons invested in it tend to find larger effect sizes in favour of their treatment than for comparison groups.¹ This finding has been extensively replicated^{2 3} and is also robust when the quality of research is controlled for. Researcher allegiance is subject of ongoing debates about the design of efficacy studies as well as implications for policy.^{2 4 5} Researcher allegiance is also discussed widely in the literature on experimental as well as evaluation research.⁶ Since the motivational underpinnings of allegiance effects are potentially far more ingrained into human behaviour and decision making than previously thought,⁷ they may occur commonly in clinical research in general.

Although it has been suggested that allegiance effects may play a role in the validation of psychological screening and case-finding tools (eg, O'Shea *et al.*, in press), systematic



evaluations of this hypothesis are rare and studies that acknowledge potential allegiance effects in such studies mainly come from forensic psychology and psychiatry backgrounds.^{8–11} Diagnostic validation studies are geared at establishing the sensitivity and specificity of a screening or case-finding tool, which is used in practice to differentiate cases from non-cases or to decide about whether further assessment or treatment is indicated or will be offered. An allegiance effect in such studies would be seen in systematically higher sensitivities or specificities if the original author(s) is (are) part of the team of such a study. Such a bias would have a deleterious affect on practice through promising overoptimistic accuracy of the screening or case-finding tool or in evaluating the cost-effectiveness of the measure in a screening or case-finding context.

The depression module of the Patient Health Questionnaire (PHQ-9) is a widely used depression-screening instrument in non-psychiatric settings. The PHQ-9 was developed by a team of researchers, with its development underwritten by an educational grant from Pfizer US Pharmaceuticals.¹² The PHQ-9 can be scored using different methods, including an algorithm based on Diagnostic and Statistical Manual of Mental Disorders (DSM)-IV criteria and a cut-off based on summed-item scores. The psychometric properties of these two approaches have been summarised in two recently published meta-analyses.^{13 14} The goal of the current review is to investigate, based on an established database of PHQ-9 diagnostic validation studies,^{13 14} whether an allegiance effect is found that leads to an increased sensitivity and specificity in studies that were conducted by researchers closely connected to the original developers of the instrument.

METHODS

Study selection

Similar search strategies were used in both systematic reviews (for full details, please see Manea *et al* and Moriarty *et al*^{13 14}). Embase, Medline and PsycINFO were searched from 1999 (when the PHQ-9 was first developed) to August 2013 and September 2013, respectively, using the terms 'PHQ-9', 'PHQ', 'PHQ\$' and 'patient health questionnaire'. The search strategy is presented in online supplementary appendix 1. The reference lists of studies fitting the inclusion criteria were manually searched and a reverse citation search in Web of Science was performed. The authors of unpublished studies were contacted and conference abstracts were reviewed in an attempt to minimise publication bias.

The following inclusion-exclusion criteria were used:

Population: adult population. *Instrument:* studies that used the PHQ-9. *Comparison (reference standard):* the accuracy of the PHQ-9 had to be assessed against a recognised gold-standard instrument for the diagnosis of either DSM or International Classification of Disease (ICD) criteria for major depression. Studies were included if the diagnoses were made using a standardised diagnostic structured

interview schedule (eg, Mini International Neuropsychiatric Interview (MINI), Structured Clinical Interview for DSM Disorders (SCID)). Unguided clinician diagnoses with no reference to a standard structured diagnostic schedule or comparisons of the PHQ-9 with other self-report measures were excluded. Studies were also excluded if the target diagnosis was not major depressive disorder (MDD, eg, any depressive disorder). *Outcome:* studies had to report sufficient information to calculate a 2×2 contingency table for the algorithm or the recommended cut-off point 10. *Study design:* any design. *Additional criterion:* we avoided double counting of evidence by ensuring that only one study of those that reported overlapping datasets in different journals were included in the meta-analysis. 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.

Quality assessment

Quality assessment was performed using the Quality Assessment of Diagnostic Accuracy Studies (Revised) (QUADAS-2) tool, a tool for evaluating the risk of bias and applicability of primary diagnostic accuracy studies when conducting diagnostic systematic reviews.¹⁵ It covers the areas of patient selection, index test, reference standard and flow and timing.¹⁶ This tool was adapted for the two reviews and quality assessments were carried out by two independent reviewers for all studies included in the reviews.

Data synthesis and statistical analysis

We constructed 2×2 tables for cut-off point 10¹⁴ and the algorithm scoring method.¹³ Pooled estimates of sensitivity, specificity, positive/negative likelihood ratios and diagnostic ORs (DOR) were calculated using random effects bivariate meta-analysis.¹⁷ Heterogeneity was assessed using I^2 for the DOR, an estimate of the proportion of study variability that is due to between-study variability rather than sampling error. We considered values of $\geq 50\%$ to indicate substantial heterogeneity.¹⁸ Summary receiver operating characteristic curves (sROC) were constructed using the bivariate model to produce a 95% confidence ellipse within ROC space.¹⁹ Each data point in the sROC space represents a separate study, unlike a traditional ROC plot, which explores the effect varying thresholds on sensitivity and specificity in a single study.

We undertook a meta-regression analysis of logit DOR using research allegiance as covariate in the meta-regression model.^{20 21} Analyses were conducted using STATA V.12, with the metan, metandi and metareg user-written commands.

Allegiance rating

We rated authorship on a paper if any of the developers of the PHQ-9—Kurt Kroenke, MD, Robert L Spitzer, MD and Janet BW Williams—as an indicator of potential allegiance. We also rated as evidence of allegiance as



acknowledged collaborations with the developers of the PHQ-9, even if they were not listed as coauthors or if the authors acknowledged funding from Pfizer to conduct the study.

RESULTS

Overview of included studies

Thirty-one studies reported the diagnostic properties of the PHQ-9 at cut-off point 10 or above and were included in this analysis.¹⁴ Twenty-seven studies were included in the algorithm review.¹³ The study selection flow charts can be found in online supplementary appendix 2 (figures 1 and 2). The characteristics of these studies are reported in [tables 1 and 2](#) and the results of the methodological assessment are presented in [tables 3 and 4](#).

Algorithm scoring method

Descriptive characteristics

The descriptive characteristics of the included studies are presented in [table 1](#). Seven individual studies that reported the diagnostic performance of the PHQ-9 using the algorithm scoring method were coauthored by the original developers of the PHQ-9,^{22–26} specifically acknowledged one of the developers and support by an educational grant from Pfizer USA,²⁷ or were coauthored by the first author of a previous study that had also been coauthored by one of the developers.²⁸ Twenty non-allegiant studies reported the diagnostic properties of the PHQ-9 using the algorithm scoring method.

Three (43%, 3/7) of the allegiant studies were conducted exclusively in hospital settings.^{22 26 28} The remaining four studies (67%, 4/7) were conducted in different settings or non-exclusively hospital settings: one in primary care²⁵ and three in mixed settings: psychosomatic walk in clinics and family practices²³,ⁱ outpatient clinics and family practices²⁴ and primary care and hospital settings.²⁷ In the non-allegiant group, 13 (65%, 13/20) studies were conducted in hospital settings.^{29–41} Of the remaining seven studies, six were conducted in primary care settings^{42–47} and one in a community sample.⁴⁸

In both groups (non-allegiant and allegiant studies), the majority of studies validated a translated version of the PHQ-9. Two of the studies authored by developers (28%, 2/7),^{25 26} and eight (40%, 8/20) allegiant studies^{29 30 37–40 42 48} were conducted in English.

The mean prevalence of MDD in the group of allegiant studies was 13.4% (range 6.1%–29.2%); in the non-allegiant group it was 15.5% (range 3.9%–32.4%). The mean age of patients in the PHQ-9 developers group was 45.7; all but one study had a mean age in the range of 40–50 years. In the non-allegiant group, the mean age was 54.6 (range

29.3–75.0), with almost half (8) of the studies reporting a mean age of over 60. The percentage of females in the PHQ-9 developers was 56.8% (range 28.6%–67.8%) and in the non-allegiant group was 59.1 (18%–100%).

All allegiant studies used a self-reported PHQ-9, whereas in seven non-allegiant studies (30%, 6/20) the PHQ-9 was administered by a researcher.^{30–33 43 48} Apart from Muramatsu *et al.*, all allegiant studies used the SCID as a gold standard²⁷; the non-allegiant studies used a wider range of gold standards including SCAN, CIDI, MINI and C-DIS, although the SCID was also frequently used by the independent studies as well (45%, 9/20 studies).

Four out of the seven allegiant studies (57%) did not include a conflict of interest statement.^{22 23 25 27} Also, four (57%) of the allegiant studies acknowledged funding from Pfizer.^{23–25 27} Only one study²⁷ acknowledged the collaboration with one of the developers of the PHQ-9.

Of the non-allegiant studies, 12 (60%) did not include a conflict of interest statement.^{29–32 35–37 39 44–46 48} It appears that newer studies were more likely to include a conflict of interest statement, which may reflect a recent change in reporting. Funding was acknowledged by most studies (18/20) and most received funding from academic or/and health research institutions. Two studies received funding from pharmaceutical companies—Lundbeck⁴³ and Pfizer,³⁵ and one study acknowledged that Pfizer Italia provided the Italian version of PHQ-9 and gave the authors permission to use it.³⁶

Diagnostic test accuracy

Pooled sensitivity and specificity was calculated separately for the non-allegiant and allegiant studies. Pooled sensitivity for the allegiant studies of the PHQ-9 was 0.77 (95% CI 0.70 to 0.84), pooled specificity was 0.94 (95% CI 0.90 to 0.97) and the pooled DOR was 64.40 (95% CI 34.15 to 121.43). Heterogeneity was high ($I^2=78.9%$). [Figure 1](#) represents the sROCs for this set of studies.

Pooled sensitivity for the non-allegiant studies was lower compared with the developer authored studies group at 0.48 (95% CI 0.41 to 0.91), pooled specificity was the same at 0.94 (95% CI 0.91 to 0.95). The pooled DOR was approximately four times lower at 15.05 (95% CI 11.03 to 20.52) (see [figure 1](#)). Heterogeneity was substantial at $I^2=68.1%$.

The meta-regression analysis for algorithm studies with non-allegiant status as the predictor of the DOR showed that non-allegiant status was a significant predictor of the DOR ($p<0.0001$) and explained a substantial amount of the observed heterogeneity (51.5%).

Quality assessment

The results of the quality assessment using QUADAS-2 are given in [table 3](#) for the studies reporting on the diagnostic performance of the algorithm scoring method. In the patient selection domain, more non-allegiant studies (65%, 13/20) than allegiant (29%, 2/7) met the criterion for consecutive referrals. There were no marked differences on the other two criteria in this domain (avoid

ⁱ This study provided separate estimates for the two settings in which it was conducted; therefore separate psychometric estimates were generated for each sample for both algorithm scoring method and summed items scoring method at cut-off point 10 (see below).

Table 1 Descriptive characteristics of algorithm studies¹³

Study	Sample characteristics			PHQ-9 characteristics	Diagnostic standard	a) COI declaration		
	(country, setting, age, sex)	Sample size and % depressed	Administration: self-report Language:			b) Funding	c) Relationship with original developers	
Diez-Quevedo <i>et al</i> ²²	Country: Spain Setting: medical and surgical tertiary hospitals Age (years): M=43 (SD=14.2) Female: 45.6%	n=1003 Depressed: 8.2%	Administration: self-report Language: Spanish	DSM-III-R SCID	a) No COI declaration b) Funding acknowledged (academic institutions) c) Not acknowledged			
Gräfe <i>et al</i> ²³	Country: Germany Setting: psychosomatic walk-in clinics and family practices Age (years): male=41.9 (SD=13.8) Female: 67.8%	n=528 Depressed: 29.2% psychosomatic patients: 6.16% medical patients	Language: German Administration: self-report	DSM-IV SCID	a) No COI declaration b) Acknowledged funding from Pfizer c) Not acknowledged			
Lowe <i>et al</i> ²⁴	Country: Germany Setting: outpatient clinics and family practices Age (years): male=41.7 (SD=13.8) Female: 67.1%	n=501 Depressed: 13.2%	Administration: self-report Language: German	DSM-IV SCID	a) COI declaration "This study was supported by unrestricted restricted grants from Pfizer Germany and from the medical faculty of the University of Heidelberg Germany, and there are no COI". b) Acknowledged funding from Pfizer and academic institution c) Not acknowledged			
Muramatsu <i>et al</i> ²⁷	Country: Japan Setting: primary care and general hospital Age (years): male=43.3 (SD=16.4) Female: 59.5%	n=131 Depressed: 28.2%	Administration: self-report Language: Japanese	DSM-IV MINI	a) No COI declaration b) Acknowledged funding from Pfizer c) Acknowledged one of the developers of the PHQ-9: 'The authors acknowledge Dr RL Spitzer'			
Navinés <i>et al</i> ²⁸	Country: Spain Setting: general hospital (patients with chronic HCV) Age (years): male=43.4 (SD=10.2) Female: 28.6%	n=500 Depressed: 6.4%	Administration: self-report Language: Spanish	DSM-IV SCID	a) All authors declared that they had no COI. b) Role of funding source declared c) Not acknowledged			
Spitzer <i>et al</i> ²⁵	Country: USA Setting: primary care Age (years): male=46 (SD=17.2) Female: 66%	n=3000 (585 received SCID) Depressed: 10%	Administration: self-report Language: English	DSM-III-R SCID	a) No COI declaration b) Acknowledged funding from Pfizer. 'Drs Spitzer and Williams receive honoraria and consulting money from Pfizer, which has supported this work'. c) N/A			
Thekkumpurath <i>et al</i> ²⁶	Country: UK Setting: hospital (cancer patients) Age (years): male=61 Female: 63%	n=782 Depressed: 6.3% (of the whole sample)	Administration: not stated Language: English	DSM-IV SCID	a) COI declaration: 'Supported by Cancer Research UK' b) As in a) c) Not acknowledged			
Ayalon <i>et al</i> ⁴³	Country: Israel Age (years): male=75 (SD=8.1) Female: 40.5%	n=153 Depressed: 3.9%	Administration: researcher administered Language: Hebrew	DSM-IV SCID	a) COI declaration: 'The project was funded by an Investigator's Initiated Research Grant from Lundbeck International given to Dr Liat Ayalon. Lundbeck International had no other involvement in the project concept of design or in this paper. Per Bech has occasionally over the past 3 years until August 2008 received funding from and has been speaker or member of advisory boards for pharmaceutical companies with an interest in the drug treatment of affective disorders (AstraZeneca, Lilly, H Lundbeck A/S, Lundbeck Foundation and Organon)'. b) Acknowledged funding from Lundbeck International			

Continued

Table 1 Continued

Study	Sample characteristics		Sample size and % depressed	PHQ-9 characteristics	Diagnostic standard	a) COI declaration		
	(country, setting, age, sex)					b) Funding	c) Relationship with original developers	
Eack <i>et al</i> ²⁹	Country: USA Setting: community mental health centres for children Age (years): male=39.20 (SD 9.63) Female: 100%	n=50 Depressed: 28%	Administration: self-report Language: English	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic/health research institutions)			
Fann <i>et al</i> ³⁰	Country: USA Setting: trauma hospital (inpatients with traumatic brain injury) Age (years): male=42 (SD=17.9) Female: 29.1%	n=135 Depressed: 16.3%	Administration: telephone-administered Language: English	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic institutions)			
Gelaye <i>et al</i> ³¹	Country: Ethiopia Setting: general hospital Age (years): 34.9 (SD=11.6) Female: 63.1%	n=363 Depressed: 12.6%	Administration: researcher-administered Language: Amharic	DSM-IV SCAN	a) No COI declaration b) Funding acknowledged (academic/health research institutions)			
Gjerdingen <i>et al</i> ⁴⁸	Country: USA Setting: community Age (years): male=29.3 Female: 100%	n=438 Depressed: 4.6%	Administration: telephone or self-report Language: English	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic/health research institutions)			
Henkel <i>et al</i> ⁴⁴	Country: Germany Setting: primary care Age (years): not reported Female: 74%	n=448 Depressed: 10%	Administration: self-report Language: German	DSM-IV CIDI	a) No COI declaration b) Funding acknowledged (academic/health research institutions)			
Hyphantis <i>et al</i> ³²	Country: Greece Setting: hospital – rheumatology patients Age (years): male=54.2 (SD=13.5) Female: 74%	n=213 Depressed: 32.4%	Administration: researcher administered Language: Greek	DSM-IV MINI	a) No COI declaration b) No funding acknowledgement			
Inagaki <i>et al</i> ³³	Country: Japan Setting: general hospital Age whole sample (years): male=73.5 (SD=12.3) Female: 59.3%	n=104 out of 511 received MINI Depressed: 7.4%	Administration: researcher administered Language: Japanese	DSM-IV MINI	a) COI declaration: 'The authors declare that they have no competing interests'. b) Funding acknowledged (academic/health research institutions)			
Khamseh <i>et al</i> ³⁴	Country: Iran Setting: diabetes clinic Age (years): male=56.17 (SD=9.60) Female: 51.9%	n=185 Depressed: 43.2%	Administration: self-report Language: Persian	DSM-IV SCID	a) COI declaration: the authors declared no competing interests b) Funding acknowledged (academic/health research institutions)			
Lamers <i>et al</i> ⁴⁵	Country: The Netherlands Setting: primary care (elderly) Age (years): male=71.4 (SD=6.90) Female: 48.2%	n=713 Depressed: 10.7%	Administration: self-report Language: Dutch	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic/health research institutions)			

Continued

Table 1 Continued

Study	Sample characteristics			PHQ-9 characteristics	Diagnostic standard	a) COI declaration		
	(country, setting, age, sex)	Sample size and % depressed	Administration: self report Language: Thai			b) Funding	c) Relationship with original developers	
Lotrakul <i>et al</i> ⁴⁶	Country: Thailand Setting: primary care Age (years): male=45.0 (SD=14.30) Female: 73.7%	n=279 Depressed: 6.8%	Administration: self report Language: Thai	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic/health research institutions)			
Persoons <i>et al</i> ³⁵	Country: Belgium Setting: hospital (otolaryngology patients) Age (years): male=48.2 (SD=12.9) Female: 65.6%	n=268 (97 received MINI) Depressed: 16.5%	Administration: self-report Language: Dutch	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic/health research institutions) and Pfizer Belgium			
Picardi <i>et al</i> ⁶⁶	Country: Italy Setting: hospital (dermatology inpatients) Age (years): male=37.5 Female: 56%	n=141 Depressed: 8.5%	Administration: self-report Language: Italian	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic/health research institutions) Acknowledged Pfizer Italia SRL for providing the Italian version of the PHQ-9 and for permission to use it.			
Stafford <i>et al</i> ³⁷	Country: Australia Setting: hospital (cardiology patients) Age (years): male=64.1 (SD=10.3) Female: 66%	n=193 Depressed: 18%	Administration: self-report Language: English	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic/health research institutions)			
Thombs <i>et al</i> ³⁸	Country: USA Setting: hospital (outpatients with coronary heart disease) Age (years): male=67 (SD=11) Female: 18%	n=1024 Depressed: 22%	Administration: not stated Language: English	DSM C-DIS	a) COI declaration 'None disclosed' b) Funding acknowledged (academic/health research institutions)			
Thompson <i>et al</i> ³⁹	Country: USA Setting: patients with Parkinson's disease Age (years): 72.5 (SD=9.6) Female: 42%	n=214 Depressed: 14%	Administration: self administered Language: English	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic/health research institutions)			
Turner <i>et al</i> ⁴⁰	Country: Australia Setting: stroke patients Age (years): 66.7 (SD=13.1) Female: 47.2%	n=72 Depressed: 18%	Administration: self administered Language: English	DSM-IV SCID	a) COI declaration: disclosures 'none'. b) Funding acknowledged (academic/health research institutions)			
van Steenberg-Weijnenburg <i>et al</i> ⁴¹	Country: The Netherlands Setting: patients with diabetes Age (years): male=61.8 (SD=13.6) Female: 48.7%	n=197 Depressed: 18.8%	Administration: self administered Language: Dutch	DSM-IV SCID	a) COI declaration: 'The authors declare that they have no competing interests'. b) Funding acknowledged (academic/health research institutions)— 'this had no influence on the content of this article'.			
Zuithoff <i>et al</i> ⁴⁷	Country: The Netherlands Setting: primary care Age (years): male=51 (SD=16.7) Female: 63%	n=1338 Depressed: 13%	Administration: self-report Language: Dutch	DSM-IV CIDI	a) COI declaration 'The authors declare that they have no competing interests'. b) Funding acknowledged (academic/health research institutions)			

CIDI, Composite International Diagnostic Interview; CIS-R, Clinical Interview Schedule; COI, conflict of interest; DSM, Diagnostic and Statistical Manual of Mental Disorders; MINI, Mini-International Neuropsychiatric Interview; N/A, not available; SCAN, Schedules for Clinical Assessments in Neuropsychiatry; SCID, Structured Clinical Interview for DSM Disorders.

Table 2 Descriptive characteristics of the summed items scoring method studies cut-off point 10¹⁴

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard	a) COI declaration b) Funding c) Relationship with original developers
13. Gräfe <i>et al</i> ²³	Country: Germany Setting: psychosomatic walk-in clinics and family practices Mean age: 41.9 (SD=13.8) Female: 67.8%	n=528 Depressed: 29.2% psychosomatic patients; 6.16% medical patients	Administration: self-report Language: German Cut-offs: 10–14	DSM-IV SCID	a) No COI declaration b) Acknowledged funding from Pfizer c) Not acknowledged
16. Kroenke <i>et al</i> ¹²	Country: USA Setting: primary care Mean age: 46 (SD=17) Female: 66%	n=580 7.1% MDD	Administration: self-report Language: English Cut-offs: 9–15	DSM-IV SCID	a) No COI declaration b) Acknowledged funding from Pfizer c) N/A
22. Navinés <i>et al</i> ²⁸	Country: Spain Setting: general hospital (patients with chronic HCV) Mean age: 43.4 (SD=10.2) Female: 28.6%	n=500 6.4% MDD	Administration: self-report Language: Spanish Cut-offs: 10	DSM-IV SCID	a) All authors declared that they had no COI b) Role of funding source declared c) Not acknowledged
29. Thekkumpurath <i>et al</i> ²⁶	Country: UK Setting: hospital (cancer patients) Mean age: 61 Female: 63%	n=782 6.3% MDD (of the whole sample)	Administration: not stated Language: English Cut-offs: 5–10	DSM-IV SCID	a) COI declaration: 'Supported by Cancer Research UK' b) As in a) c) Not acknowledged
33. Williams <i>et al</i> ⁴⁹	Country: USA Setting: secondary care (poststroke) Mean age: unclear Female: unclear	n=316 33.5% MDD	Administration: unclear Language: English Cut-offs: 10	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic institutions) c) Not acknowledged
1. Adewuya <i>et al</i> ⁵⁵	Country: Nigeria Setting: community (students) Mean age: 24.8 (15–40) Female: 41.2%	n=512 2.5% MDD	Administration: Self-report Language: English Cut-offs: 8–12	DSM-IV MINI	a) No COI declaration b) No funding declaration
2. Arroll <i>et al</i> ⁴²	Country: New Zealand Setting: primary care Mean age: 49 (17–99) Female: 61%	n=2642 6.2% MDD	Administration: not stated Language: English Cut-offs: 8, 10, 12, 15	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
3. Azah <i>et al</i> ⁶²	Country: Malaysia Setting: primary care Mean age: 38.7 (18–79) Female: 61.7%	n=180 16.6% MDD	Administration: self-report Language: Malay Cut-offs: 5–12	DSM-IV CIDI	b) No COI declaration c) Funding acknowledged (academic/health research institutions)
4. Chagas <i>et al</i> ⁵⁰	Country: Brazil Setting: secondary care Mean age: not stated Female: 52.7%	n=84 25.5% MDD	Administration: self-report Language: Brazilian Cut-offs: 7–10	DSM-IV SCID	a) COI declaration 'None declared' b) Funding acknowledged (academic/health research institutions)

Continued



Table 2 Continued

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard	a) COI declaration b) Funding c) Relationship with original developers
6. de Lima Osorio <i>et al</i> ⁶⁰	Country: Brazil Setting: primary care Mean age: unclear Female: 100%	n=177 34% MDD	Administration: research assistants Language: Brazilian Portuguese Cut-offs: 10–15	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic institutions)
7. Elderon <i>et al</i> ⁵¹	Country: USA Setting: secondary care Mean age: unclear Female: 18%	n=1022 18.3% MDD	Administration: self-report Language: English Cut-offs: 10	C-DIS	a) COI declaration—'No disclosures' b) Funding acknowledged (academic institutions and industry—AHA Pharmaceuticals Roundtable)—'The funding organisations had no role in the design or conduct of the study, collection, management, analysis or interpretation of data; or preparation, review or approval of the manuscript'.
8. Fann <i>et al</i> ⁶⁰	Country: USA Setting: trauma hospital (inpatients with traumatic brain injury) Mean age: 42 (SD=17.9) Female: 29.1%	n=135 16.3% MDD	Administration: telephone-administered Language: English Cut-offs: 10	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic institutions)
9. Fine <i>et al</i> ⁵⁶	Country: USA Setting: primary care (Ohio Army National Guard) Mean age: 31 (17–60) Female: 12%	n=498 21.5% MDD	Administration: telephone-administered Language: English Cut-offs: 10, 15	DSM-IV SCID-I	a) COI—last author disclosed financial and consulting interests (Pfizer not one of them). All other authors declared that they have no COI. b) Funding acknowledged—DoD Medical Research. 'The sponsor had no role in study design, data collection, analysis, interpretation of results, report writing or manuscript submission'.
10. Gelaye <i>et al</i> ³¹	Country: Ethiopia Setting: general hospital Mean age: 34.9 (SD=11.6) Female: 63.1%	n=363 12.6% MDD	Administration: researcher-administered Language: Amharic Cut-offs: 9–11	DSM-IV SCAN	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
11. Gilbody <i>et al</i> ⁵⁷	Country: UK Setting: primary care Mean age: 42.5 (SD 13.6) Female: 77%	n=96 37.5 MDD	Administration: not stated Language: English Cut-offs: 9–13	DSM-IV SCID	a) COI declaration—last author involved in the development of one of the instruments (CORE-OM), 'but does not gain financially from its use.' b) Funding acknowledged (academic/health research institutions)
12. Gierdingen <i>et al</i> ⁴⁸	Country: USA Setting: community Mean age: 29.3 Female: 100%	n=438 4.6% MDD	Administration: telephone or self-report Language: English Cut-offs: 10	DSM-IV SCID	a) No COI declaration c) Funding acknowledged (academic/health research institutions)
14. Hyphantis <i>et al</i> ⁶²	Country: Greece Setting: hospital—rheumatology patients Mean age: 54.2 (SD=13.5) Female: 74%	n=213 32.4% MDD	Administration: researcher administered Language: Greek Cut-offs: 4–16	DSM-IV MINI	a) No COI declaration b) No funding acknowledgement

Continued

Table 2 Continued

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard	a) COI declaration b) Funding c) Relationship with original developers
15. Khamseh et al ⁶⁴	Country: Iran Setting: outpatient diabetic clinic Mean age: 56.1 (SD=9.6) Female: 51.8%	n=185 43.2% MDD	Administration: self-report Language: Persian Cut-offs: 10, 13	DSM-IV SCID	a) COI declaration: the authors declared no competing interests. b) Funding acknowledged (academic/health research institutions)
19. Liu et al ⁶³	Country: Taiwan Setting: primary care Mean age: not specified Female: 60.9%	n=1532 3.3% MDD	Administration: self-report Language: Chinese version Cut-offs: 9–11	SCAN	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
20. Lotrakul et al ⁴⁶	Country: Thailand Setting: primary care Mean age: 45.0 (SD=14.30) Female: 73.7%	n=279 6.8% MDD	Administration: self report Language: Thai Cut-offs: 7–15	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
23. Patel et al ⁶¹	Country: India Setting: primary care Mean age: 37.5 (18–83) Female: 56.4%	n=299 4.3% MDD	Administration: face-to-face interview Language: not specified Cut-offs: 7–15	CIS-R	a) COI declaration—No declaration of interest b) Funding acknowledged (academic/health research institutions)
24. Phelan et al ⁶⁸	Country: USA Setting: primary care (elderly) Mean age: 78 (SD=7) Female: 62%	n=71 12% MDD	Administration: research assistant Language: English Cut-offs: 8–12	DSM-IV SCID	a) COI declaration—no competing interests b) Funding acknowledged (academic/health research institutions). 'The funder had no role in the study design, methods, data collection, analysis or interpretation of data, nor any role in the preparation of the manuscript or decision to submit the manuscript for publication'.
25. Rooney et al ⁶²	Country: UK Setting: secondary care (glioma) Mean age: 54.2 (SD=12.3) Female: 42.6%	n=129 13.5% MDD	Administration: self-report Language: English Cut-offs: 8–11	DSM-IV SCID	a) COI declaration 'The authors declare that they have no COI'. b) Funding acknowledged (academic/health research institutions)
26. Sherina et al	Country: Malaysia Setting: primary care Mean age: 30.9 (18–81) Female: 100%	n=146 21.2% MDD	Administration: self-report Language: Malay Cut-offs: 10	CIDI	a) COI declaration 'The authors declare that they have no competing interests'. b) Funding acknowledged (academic/health research institutions)
27. Sidebottom et al ⁶⁹	Country: USA Setting: community (prenatal) Mean age: 23 (SD=5.5) Female: 100%	n=745 3.6% MDD	Administration: interview Language: English Cut-offs: 10	DSM-IV SCID	a) COI declaration 'The authors declare that they have no financial COI'. b) Funding acknowledged (academic/health research institutions)
28. Stafford et al ⁶⁷	Country: Australia Setting: secondary care (cardiac procedures) Mean age: 64.14 (38–91) Female: 19.2%	n=193 18.1% MDD	Administration: self-report Language: English Cut-offs: 10	DSM-IV MINI	a) No COI declaration b) Funding acknowledged (academic/health research institutions)

Continued



Table 2 Continued

Study	Sample characteristics	Sample size and % MDD	PHQ-9 characteristics	Diagnostic standard	a) COI declaration b) Funding c) Relationship with original developers
30. Thombs <i>et al</i> ³⁸	Country: USA Setting: hospital (outpatients with coronary heart disease) Mean age: 67 (SD=11) Female: 18%	n=1024 22% MDD	Administration: not stated Language: English Cut-offs: 7–10	DSM C-DIS	a) COI declaration 'None disclosed' b) Funding acknowledged (academic/health research institutions)
32. Watnick <i>et al</i> ⁵³	Country: USA Setting: secondary care (dialysis) Mean age: 63 (SD=15) Female: 32.3%	n=62 19% MDD	Administration: self-report Language: English Cut-offs: 10	DSM-IV SCID	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
34. Wittkamp <i>et al</i> ⁶⁴	Country: The Netherlands Setting: primary care Mean age: 49.8 Female: 66.7%	n=664 12.3% MDD	Administration: self-report Language: not specified Cut-offs: 10 and 15	DSM-IV SCIDI	a) No COI declaration b) Funding acknowledged (academic/health research institutions)
35. Zhang <i>et al</i> ⁵⁴	Country: Hong Kong Setting: secondary care (diabetic outpatients) Mean age: 55.1 (SD=9.5) Female: 40.8%	n=99 23.2% MDD	Administration: self-report Language: Chinese version Cut-offs: 15	DSM-IV MINI	a) COI declaration – last author acknowledged financial COI. The other authors declare that they have no competing interests. b) Funding acknowledged (academic/health research institutions)
36. Zuihoff <i>et al</i> ⁴⁷	Country: The Netherlands Setting: primary care Age (years): male=51 (SD=16.7) Female: 63%	n=1338 Depressed: 13%	Administration: self-report Language: Dutch	DSM-IV CIDI	a) COI declaration 'The authors declare that they have no competing interests'. b) Funding acknowledged (academic/health research institutions)

COI, conflict of interest; DSM, Diagnostic and Statistical Manual of Mental Disorders; MDD, major depressive disorder; N/A, not available; SCID, Structured Clinical Interview for DSM Disorders.

Table 3 Quality assessment of included studies in the algorithm meta-analysis¹³

Study	Patient selection:		Patient selection:		Patient selection:	Index test:			Index test:	Index test:
	Consecutive or random sample	Avoid case-control/avoid artificially inflated base rate	Avoided inappropriate exclusions	Overall risk of bias		PHQ-9 interpreted blind to reference test	If translated, appropriate translation	If translated, psychometric properties reported		
Allegiant studies										
Diez-Quevedo <i>et al</i> ²²	X	✓	X	High	?	✓	✓	✓	Unclear	Unclear
Gräfe <i>et al</i> ²³	✓	✓	✓	Low	?	✓	✓	✓	Unclear	Unclear
Lowe <i>et al</i> ²⁴	X	✓	✓	High	✓	✓	✓	✓	Low	Low
Muramatsu <i>et al</i> ²⁷	?	✓	?	Unclear	✓	✓	?	?	Unclear	Unclear
Navines <i>et al</i> ²⁸	✓	✓	✓	Low	✓	✓	?	?	Unclear	Unclear
Spitzer <i>et al</i> ²⁵	X	✓	✓	High	✓	N/A	N/A	N/A	Low	Low
Thekkumpurath <i>et al</i> ²⁶	X	X	✓	High	✓	N/A	N/A	N/A	Low	Low
Non-allegiant studies										
Arroll <i>et al</i> ⁴²	✓	✓	✓	Low	✓	N/A	N/A	N/A	Low	Low
Ayalon <i>et al</i> ⁴³	?	✓	✓	Unclear	?	✓	?	?	Unclear	Unclear
Eack <i>et al</i> ²⁹	?	✓	?	Unclear	?	N/A	N/A	N/A	Unclear	Unclear
Fann <i>et al</i> ³⁰	✓	X	X	High	✓	N/A	N/A	N/A	Low	Low
Gelaye <i>et al</i> ³¹	?	X	?	High	✓	✓	?	?	Unclear	Unclear
Gjerdengen <i>et al</i> ⁴⁸	✓	✓	✓	Low	?	N/A	N/A	N/A	Unclear	Unclear
Henkel <i>et al</i> ⁴⁴	✓	✓	✓	Low	?	N/A	N/A	N/A	Unclear	Unclear
Hyphantis <i>et al</i> ³²	✓	✓	X	High	✓	?	?	?	Unclear	Unclear
Inagaki <i>et al</i> ³³	✓	X	✓	High	✓	?	?	?	Unclear	Unclear
Khamseh <i>et al</i> ³⁴	✓	✓	?	Unclear	✓	✓	?	?	Unclear	Unclear
Lamers <i>et al</i> ⁴⁵	✓	X	X	High	✓	?	?	?	Unclear	Unclear
Lotrakul <i>et al</i> ⁴⁶	X	✓	?	High	✓	✓	?	?	Unclear	Unclear
Persoons <i>et al</i> ³⁵	✓	✓	✓	Low	✓	✓	✓	N/A	Low	Low
Picardi <i>et al</i> ³⁶	✓	✓	✓	Low	✓	?	?	?	Unclear	Unclear
Stafford <i>et al</i> ³⁷	✓	✓	✓	Low	✓	N/A	N/A	N/A	Low	Low
Thombs <i>et al</i> ³⁸	X	✓	?	Unclear	?	N/A	N/A	N/A	Unclear	Unclear
Thomson <i>et al</i> ³⁹	?	✓	✓	Unclear	?	N/A	N/A	N/A	Unclear	Unclear
Turner <i>et al</i> ⁴⁰	✓	✓	✓	Low	✓	N/A	N/A	N/A	Low	Low

Continued



Table 3 Continued

Study	Patient selection:		Patient selection:		Patient selection:		Index test:		Index test:		Index test:	
	Consecutive or random sample	avoid case-control/artificially inflated base rate	Avoided inappropriate exclusions	Overall risk of bias	Overall risk of bias	Overall risk of bias	PHQ-9 interpreted blind to reference test	If translated, appropriate translation	If translated, psychometric properties reported	Overall risk of bias	Overall risk of bias	Overall risk of bias
van Steenberg-Wijnenburg <i>et al</i> ⁴¹	?	✓	✓	Unclear	Unclear	?	?	?	?	?	Unclear	Unclear
Zuithoff <i>et al</i> ⁴⁷	✓	✓	✓	Low	Low	✓	✓	✓	?	?	Unclear	Unclear
	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:	Reference test:
	Reference test correctly classifies target condition	Reference test interpreted blind to PHQ-9	Reference test appropriate translation	Reference test psychometric properties reported	Reference test Overall risk of bias	Reference test Overall risk of bias	Reference test Interval of 2 weeks or less	Reference test All participants receive same reference test	Reference test All participants included in analysis?	Reference test Overall risk of bias	Reference test Overall risk of bias	Reference test Overall risk of bias
	Allegiant studies											
Diez-Quevedo <i>et al</i> ²²	✓	✓	✓	?	Unclear	Unclear	✓	✓	✓	✓	Low	Low
Gräfe <i>et al</i> ²³	✓	?	N/A	N/A	Unclear	Unclear	✓	✓	✓	✓	Low	Low
Lowe <i>et al</i> ²⁴	✓	✓	N/A	N/A	Low	Low	✓	✓	✓	✓	Low	Low
Muramatsu <i>et al</i> ²⁷	✓	✓	✓	✓	Low	Low	✓	✓	?	?	Unclear	Unclear
Navines <i>et al</i> ²⁸	✓	✓	?	?	Unclear	Unclear	✓	✓	✓	✓	Low	Low
Spitzer <i>et al</i> ²⁵	✓	✓	N/A	N/A	Low	Low	✓	✓	✓	✓	High	High
Thekkumpurath <i>et al</i> ²⁶	✓	✓	N/A	N/A	Low	Low	?	✓	✓	✓	High	High
	Non-allegiant studies											
Arroll <i>et al</i> ⁴²	✓	✓	N/A	N/A	Low	Low	✓	✓	✓	✓	Low	Low
Ayalon <i>et al</i> ⁴³	✓	?	✓	?	Unclear	Unclear	?	✓	✓	✓	Unclear	Unclear
Eack <i>et al</i> ²⁹	✓	?	N/A	N/A	Unclear	Unclear	?	✓	?	?	Unclear	Unclear
Fann <i>et al</i> ³⁰	✓	?	N/A	N/A	Unclear	Unclear	✓	✓	✓	✓	High	High
Gelaye <i>et al</i> ³¹	✓	✓	✓	✓	Low	Low	✓	✓	✓	✓	High	High
Gjerdengen <i>et al</i> ⁴⁸	✓	?	N/A	N/A	Unclear	Unclear	✓	✓	✓	✓	High	High
Henkel <i>et al</i> ⁴⁴	✓	?	N/A	N/A	Unclear	Unclear	✓	✓	✓	✓	High	High
Hyphantis <i>et al</i> ³²	✓	✓	?	?	Unclear	Unclear	✓	✓	✓	✓	High	High
Inagaki <i>et al</i> ³³	✓	✓	✓	?	Unclear	Unclear	✓	✓	✓	✓	High	High
Khamseh <i>et al</i> ³⁴	✓	✓	✓	?	Unclear	Unclear	✓	✓	?	?	Unclear	Unclear
Lamers <i>et al</i> ⁴⁵	✓	✓	?	?	Unclear	Unclear	?	✓	✓	✓	High	High

Continued

**Table 4** Quality assessment of included studies in the summed item scoring method cut-off point 10 meta-analysis¹⁴

Study	Patient selection:		Patient selection:		Patient selection:		Patient selection:		Patient selection:		Patient selection:		Overall risk of bias	Overall risk of bias	Overall risk of bias	Overall risk of bias	Overall risk of bias
	Consecutive or random sample	Avoid case-control/artificially inflated base rate	Avoided inappropriate exclusions	Overall risk of bias	PHQ-9 interpreted blind to reference test	Was a threshold prespecified?	If translated, appropriate translation	If translated, psychometric properties reported	Index test:	Index test:	Index test:	Index test:					
Allegiant studies																	
13. Gräfe <i>et al</i> ²³	✓	✓	✓	Low	?	✓	✓	✓	✓	✓	✓	✓	Unclear	Unclear	Unclear	Unclear	Unclear
16. Kroenke <i>et al</i> ¹²	✓	✓	✓	Low	✓	✓	✓	N/A	N/A	N/A	N/A	N/A	Low	Low	Low	Low	Low
22. Navinés <i>et al</i> ²⁸	✓	✓	✓	Low	✓	✓	✓	✓	✓	✓	✓	?	Unclear	Unclear	Unclear	Unclear	Unclear
29. Thekkumpurath <i>et al</i> ²⁶	×	×	✓	High	✓	✓	✓	N/A	N/A	N/A	N/A	N/A	Low	Low	Low	Low	Low
33. Williams <i>et al</i> ⁴⁹	✓	✓	✓	Low	?	✓	✓	N/A	N/A	N/A	N/A	N/A	Unclear	Unclear	Unclear	Unclear	Unclear
Non-allegiant studies																	
1. Adewuya <i>et al</i> ⁵⁵	✓	✓	×	Unclear	✓	✓	✓	N/A	N/A	N/A	N/A	N/A	Low	Low	Low	Low	Low
2. Arroll <i>et al</i> ⁴²	✓	✓	✓	Low	✓	✓	✓	N/A	N/A	N/A	N/A	N/A	Low	Low	Low	Low	Low
3. Azah <i>et al</i> ⁶²	✓	×	?	High	✓	✓	✓	✓	✓	✓	✓	✓	Low	Low	Low	Low	Low
4. Chagas <i>et al</i> ⁵⁰	✓	✓	✓	Low	✓	✓	✓	✓	✓	✓	✓	✓	Low	Low	Low	Low	Low
6. de Lima Osorio <i>et al</i> ⁶⁰	✓	×	✓	High	?	×	×	N/A	N/A	N/A	N/A	N/A	High	High	High	High	High
7. Elderon <i>et al</i> ⁵¹	✓	✓	✓	Low	✓	✓	✓	N/A	N/A	N/A	N/A	N/A	Low	Low	Low	Low	Low
8. Fann <i>et al</i> ⁶⁰	✓	×	×	High	✓	✓	✓	N/A	N/A	N/A	N/A	N/A	Low	Low	Low	Low	Low
9. Fine <i>et al</i> ⁵⁶	✓	✓	✓	Low	?	✓	✓	N/A	N/A	N/A	N/A	N/A	Unclear	Unclear	Unclear	Unclear	Unclear
10. Gelaye <i>et al</i> ⁶¹	?	×	?	High	✓	×	×	✓	✓	×	×	?	High	High	High	High	High
11. Gilbody <i>et al</i> ⁶⁷	?	✓	?	Unclear	✓	✓	✓	N/A	N/A	N/A	N/A	N/A	Low	Low	Low	Low	Low
12. Gierdingen <i>et al</i> ⁴⁸	✓	✓	✓	Low	?	✓	✓	N/A	N/A	N/A	N/A	N/A	Unclear	Unclear	Unclear	Unclear	Unclear
14. Hyphantis <i>et al</i> ⁶²	✓	×	✓	High	✓	✓	✓	?	?	?	?	?	Unclear	Unclear	Unclear	Unclear	Unclear
15. Khamseh <i>et al</i> ³⁴	✓	✓	?	Unclear	✓	✓	✓	✓	✓	✓	✓	?	Unclear	Unclear	Unclear	Unclear	Unclear
19. Liu <i>et al</i> ⁶³	✓	✓	?	Unclear	✓	?	?	✓	✓	×	×	?	High	High	High	High	High
20. Lotrakul <i>et al</i> ⁴⁶	×	✓	?	Unclear	✓	?	?	✓	✓	✓	✓	?	Unclear	Unclear	Unclear	Unclear	Unclear
23. Patel <i>et al</i> ⁶¹	✓	✓	✓	Low	✓	✓	✓	?	?	✓	?	?	Unclear	Unclear	Unclear	Unclear	Unclear
24. Phelan <i>et al</i> ⁶⁸	×	✓	✓	High	✓	✓	×	N/A	N/A	N/A	N/A	N/A	High	High	High	High	High
25. Rooney <i>et al</i> ⁶²	✓	✓	✓	Low	?	×	×	N/A	N/A	N/A	N/A	N/A	Low	Low	Low	Low	Low
26. Sherina <i>et al</i>	✓	✓	×	High	✓	✓	✓	✓	✓	✓	✓	✓	High	High	High	High	High
27. Sidebottom <i>et al</i> ⁶⁹	✓	✓	✓	Low	✓	✓	✓	N/A	N/A	N/A	N/A	N/A	Low	Low	Low	Low	Low
28. Stafford <i>et al</i> ³⁷	✓	✓	✓	Low	✓	✓	✓	N/A	N/A	N/A	N/A	N/A	Low	Low	Low	Low	Low

Continued



Table 4 Continued

Study	Reference test:			Reference test:		Reference test:		Reference test:		Reference test:		Reference test:		Reference test:		Reference test:		
	Reference test correctly classifies target condition	Reference test interpreted blind to PHQ-9	Reference test appropriate translation	If translated, psychometric properties reported	Overall risk of bias	Interval of 2 weeks or less	All participants receive same reference test	All participants included in analysis?	Flow / timing:	Flow / timing:	Flow / timing:	Flow / timing:	Flow / timing:	Flow / timing:	Flow / timing:	Flow / timing:	Flow / timing:	Flow / timing:
14. Hyphantis <i>et al</i> ⁶²	✓	✓	?	?	Unclear	✓	✓	×	✓	✓	×	High						
15. Khamseh <i>et al</i> ⁶⁴	✓	✓	✓	?	Unclear	✓	✓	?	✓	✓	?	Unclear						
19. Liu <i>et al</i> ⁶³	✓	✓	✓	✓	Low	✓	✓	?	✓	✓	?	Unclear						
20. Lotrakul <i>et al</i> ⁴⁶	✓	✓	✓	✓	Low	?	✓	×	✓	✓	×	High						
23. Patel <i>et al</i> ⁶¹	✓	✓	✓	?	Unclear	?	✓	×	✓	✓	×	High						
24. Phelan <i>et al</i> ⁵⁸	✓	✓	N/A	N/A	Low	✓	✓	✓	✓	✓	✓	Low						
25. Rooney <i>et al</i> ⁶²	✓	?	N/A	N/A	Unclear	?	✓	×	✓	✓	×	High						
26. Sherina <i>et al</i>	✓	✓	✓	✓	Low	✓	✓	✓	✓	✓	✓	Low						
27. Sidebottom <i>et al</i> ⁵⁹	✓	✓	N/A	N/A	Low	✓	✓	×	✓	✓	×	High						
28. Stafford <i>et al</i> ⁶⁷	✓	✓	N/A	N/A	Low	✓	✓	×	✓	✓	×	High						
30. Thombs <i>et al</i> ⁶⁸	?	✓	N/A	N/A	Unclear	✓	✓	✓	✓	✓	✓	Low						
32. Watnick <i>et al</i> ⁵³	✓	✓	N/A	N/A	Low	✓	✓	✓	✓	✓	✓	Low						
34. Wittkamp <i>et al</i> ⁶⁴	✓	✓	N/A	N/A	Low	?	✓	×	✓	✓	×	High						
35. Zhang <i>et al</i> ⁵⁴	✓	?	✓	✓	Unclear	×	✓	×	✓	✓	×	High						
36. Zuihoff <i>et al</i> ⁴⁷	✓	✓	?	?	Unclear	?	✓	✓	✓	✓	✓	Unclear						

N/A, not applicable; PHQ-9, Patient Health Questionnaire-9; ✓, criterion met; ✗, criterion not met; ?, insufficient information to code whether criterion met.

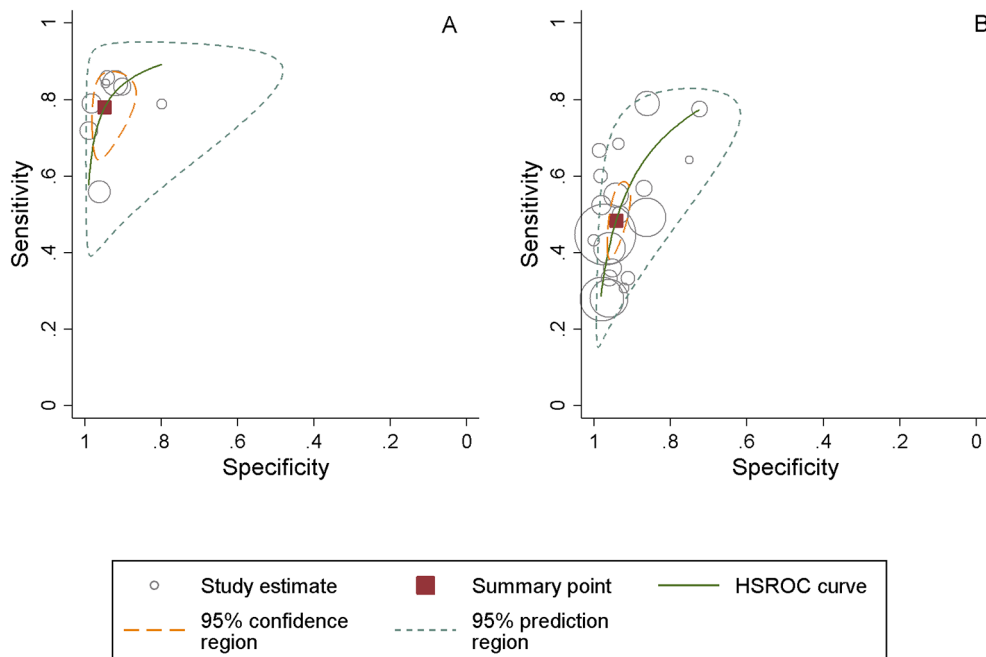


Figure 1 Patient Health Questionnaire-9 algorithm scoring method summary receiver operating characteristic plot for the diagnosis of major depressive disorder in allegiant studies (panel A) and non-allegiant studies (panel B). Pooled sensitivity and specificity estimates using a bivariate meta-analysis. HSROC, hierarchical receiver operating characteristic.

case-control design, avoid inappropriate exclusions). In the index test domain, the proportion of studies reporting that the PHQ-9 was conducted blind to the reference test was comparable between the two groups. There were differences in this domain for those studies using a translated version of the test. All non-English allegiant studies (5/5) used an appropriately translated version of the PHQ-9, whereas just over a half of the non-allegiant studies reported this (55%, 6/11). However, the majority of both sets of studies did not report details of psychometric properties of the translated version. For the reference test domain, nearly all studies in both groups were rated as using a reference test that would correctly classify the condition. While most allegiant studies reported that the reference test was interpreted blind to the PHQ-9 score (86%, 6/7), this was reported in only 60% (12/20) of the non-allegiant studies.

The two sets of studies that used translated versions of the reference test were broadly comparable. There was a slight indication that the allegiant studies were more likely to use an appropriately translated version of the reference test and report data on the psychometric properties of the translated version, although the numbers for the translated comparison are very low. There were, however, some more notable differences on the flow and timing domain. Most allegiant studies ensured that the time between the index and reference test was under 2 weeks (86%, 6/7) in comparison to 70% (14/20) of the non-allegiant studies. More allegiant studies met the criterion for ‘all participants included in the analysis’ (57%, 4/7) than non-allegiant studies (25%).

Summed items scoring method (cut-off point 10 or above)

Descriptive characteristics

Table 2 presents the sample characteristics of the 31 PHQ-9 validation studies that reported the psychometric properties of the PHQ-9 at cut-off point 10 or above. Five of these studies were coauthored by the original developers of the instrument or acknowledged collaboration^{12 23 26 49} or were coauthored by the first author of a previous study that had also been coauthored by one of the developers.²⁸ Twenty-six studies were conducted by independent researchers.

Three (60%, 3/5) allegiant studies^{26 28 49} and 11 non-allegiant studies (42%, 11/26)^{30–32 34 37 38 50–54} were conducted in hospital settings.

Three (60%, 3/5) allegiant studies^{12 26 49} and 13 non-allegiant studies (13/26)^{30 37 38 42 48 51–53 55–59} were conducted in English.

The mean prevalence of MDD in the allegiant group was 13.2% (range 6.1%–33.5%) and in the non-allegiant group was 16.1% (range 2.5%–43.2%). The mean age of patients in the allegiant group studies was 48.1 (range 41.9–61.0) and in the 26 non-allegiant studies that reported these data was 49.1 (range 23.0–78.0). The percentage of females in the allegiant studies that reported these data^{12 23 26 28} was 56.3% (range 28.6%–67.8%) and in the non-allegiant group was 64.9% (range 12%–100%).

Three allegiant studies used the self-reported mode of administration and two of them did not specify how the PHQ-9 was administered. In nine non-allegiant studies (34%, 9/26), the PHQ-9 was administered by the researcher.^{30–32 48 56 58–61} All allegiant studies used SCID as



a gold standard; the non-allegiant studies used a wider range of gold standards including SCAN, CIDI, MINI, CIS-R, C-DIS, although the SCID was used in half of the studies (50%, 13/26 studies).

Three allegiant studies (60%) did not include a conflict of interest statement.^{12 23 49} Two of these studies^{12 23} acknowledged funding from Pfizer. None of the allegiant studies acknowledged collaboration or authorship of one of the developers of the PHQ-9.

Of the non-allegiant studies, 13 (42%) did not include a conflict of interest statement.^{30–32 37 42 46 48 53 55 60 62–64} Similar to the algorithm studies, the newer studies were more likely to include a conflict of interest statement. Funding was acknowledged by most studies (27/31) and most received funding from academic and/or health research institutions. One study⁵⁷ acknowledged that the last author involved in the development of one of the instruments (CORE-OM), ‘but does not gain financially from its use’. One study⁵¹ acknowledged funding from industry, AHA Pharmaceuticals Roundtable, but stated that ‘the funding organisations had no role in the design or conduct of the study, collection, management, analysis or interpretation of data; or preparation, review or approval of the manuscript. Fine *et al.* disclosed that the last author had financial and consulting interests (Pfizer was not cited as one of them).⁵⁶

Diagnostic test accuracy

Pooled sensitivity of allegiant studies was 0.87 (95% CI 0.77 to 0.93), pooled specificity was 0.87 (95% CI 0.76 to 0.94) and the pooled DOR was 49.31 (95% CI 25.74 to 94.48)—see [table 5](#). Heterogeneity was moderate ($I^2=55.1%$). [Figure 2](#) represents the sROCs for this group.

Pooled sensitivity of non-allegiant studies was 0.76 (95% CI 0.67 to 0.83), pooled specificity was 0.88 (95% CI 0.85 to 0.91) and the pooled DOR was 24.96 (95% CI 14.81 to 42.08), approximately half that of the allegiant studies ([table 2](#)). Heterogeneity was high at $I^2=81.5%$. [Figure 2](#) represents the sROCs for this group.

The meta-regression for the studies using a cut-off point of 10 or above with allegiance status of the predictor showed that allegiance status was a significant predictor of the DOR ($p=0.015$) and explained 19.0% of observed heterogeneity.

Quality assessment

The results of the quality assessment using the QUADAS-2 are given in [table 4](#). For the patient selection domain, the two groups of studies were broadly comparable on two items (consecutive or random sample, avoid case-control design). However, all allegiant studies were rated as avoiding inappropriate exclusions (5/5) in contrast to 58% (15/26) of the non-allegiant studies.

On the index test domain, there were a number of differences between the two groups of studies. More of the non-allegiant studies (81%, 21/26) reported that the PHQ-9 was interpreted blind to the reference test compared with 60% (3/5) of the allegiant studies. All

Table 5 Pooled estimates of diagnostic properties of the Patient Health Questionnaire-9 at cut-off point 10 and using algorithm scoring method in the non-independent vs independent studies groups

Settings	No. of studies	No. of patients	Sensitivity (95% CI)	Specificity (95% CI)	Pooled positive likelihood ratio (95% CI)	Pooled negative likelihood ratio (95% CI)	Diagnostic OR (95% CI)	Heterogeneity: I^2
Manea <i>et al.</i> , 2014 SR-RA group	7	4065	0.77 (0.70 to 0.84)	0.94 (0.90 to 0.97)	14.97 (8.39 to 26.71)	0.23 (0.17 to 0.31)	64.40 (34.15 to 121.43)	78.9%
Manea <i>et al.</i> , 2014 SR Independent studies	21	9900	0.48 (0.41 to 0.91)	0.94 (0.91 to 0.95)	8.26 (6.15 to 11.09)	0.54 (0.48 to 0.62)	15.05 (11.03 to 20.52)	68.1%
Moriarty <i>et al.</i> , 2015 SR-RA group	5	6188	0.87 (0.77 to 0.93)	0.87 (0.76 to 0.94)	7.24 (3.74 to 14.03)	0.14 (0.08 to 0.25)	49.31 (25.74 to 94.48)	55.1%
Moriarty <i>et al.</i> , 2015 SR Independent studies	26	13164	0.76 (0.67 to 0.83)	0.88 (0.85 to 0.91)	6.72 (5.06 to 8.92)	0.26 (0.19 to 0.37)	24.96 (14.81 to 42.08)	81.5%

SR, Systematic review; RA, researcher allegiance.

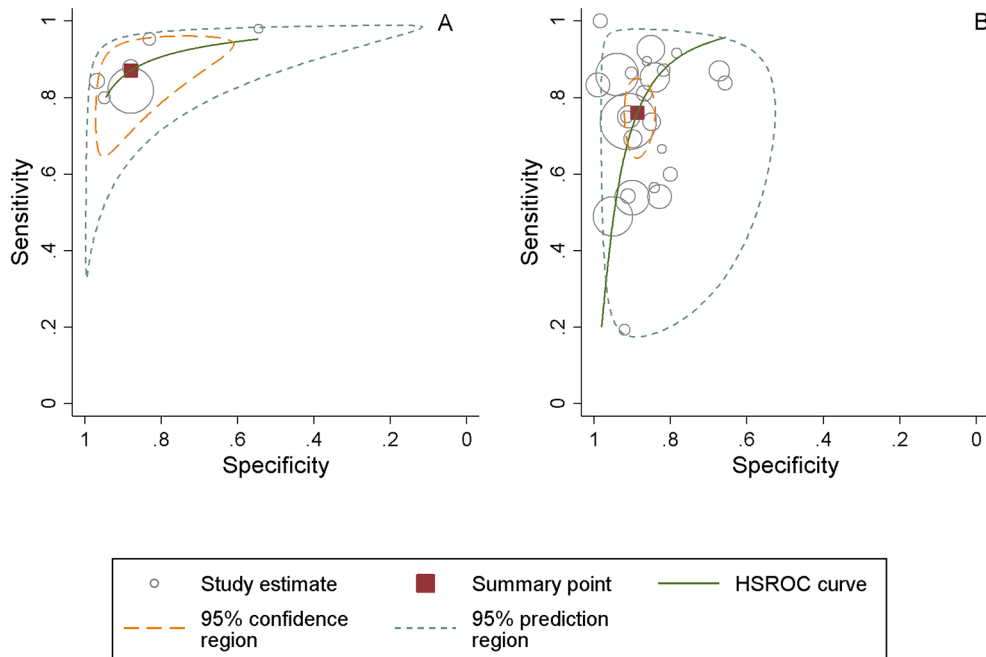


Figure 2 Patient Health Questionnaire-9 summed items scoring method at cut-off point 10 summary receiver operating characteristic plot for diagnosis of major depressive disorder in allegiant studies (panel A) and non-allegiant studies (panel B). Pooled sensitivity and specificity using a bivariate meta-analysis. HSROC, hierarchical receiver operating characteristic.

(5/5) allegiant studies were rated as prespecifying the threshold on the PHQ-9 compared with 73% (19/26) of the non-allegiant studies. The two sets of studies were broadly comparable in terms of two items from the reference test domain (correctly classify target condition, reference test interpreted blind). Only one allegiant study used a translated version of the index test or reference test, so it is not possible to comment on differences between the two sets of studies in terms of these items from the index or reference test domains. For the flow and timing domain, the two groups of studies were broadly comparable for two of the criteria (interval of 2 weeks or less, all participants receive same reference test). However, fewer than half of the non-allegiant studies met the criterion for 'all participants included in the analysis' (42%, 11/26), whereas all allegiant studies met this criterion.

DISCUSSION

This is to our knowledge the first systematic examination of a possible 'allegiance' or authorship effect in the validation of screening or case-finding psychological instrument for a common mental health disorder. We reviewed diagnostic validation studies of the PHQ-9, a widely used depression screening instrument. We found that allegiant studies reported higher sensitivity paired with similar specificity compared with non-allegiant studies. When entered as a covariate in meta-regression analyses, allegiance status was predictive of variation in the DOR for both the algorithm scoring method and the summed-item scoring method at a cut-off point of 10 or above.

Previous research has proposed several possible explanations for the allegiance effect.⁹⁻¹¹ One possibility is the

advertent bias that may serve to inflate the performance of a test when evaluated by those who have developed it. However, before concluding that the differences are due to this, it is important to explore and rule out alternative explanations. First, it is possible that any observed differences are a result of differences in study characteristics of the two sets of studies (eg, setting, clinical population). Second, differences in the methodological quality of the studies may also account for any differences. These possibilities are examined below.

Difference in study characteristics as potential alternative explanations

The two sets of studies were broadly comparable in terms of gender and the prevalence of depression, so these variables are unlikely to offer an explanation for the differences. While there were some indications from both sets of comparisons that the PHQ-9 may have been researcher-administered more often in the independent studies, it is not immediately clear how this would lead to lowered diagnostic performance.

The diagnostic meta-analyses of the PHQ-9^{13 14} have shown that the sensitivity and DOR of the PHQ-9 tends to be lower in hospital settings for both algorithm and summed-item scoring methods. While the fact that proportionally more non-allegiant algorithm studies were conducted in secondary care could explain the lower sensitivity and DOR values in the algorithm studies, in the studies that reported the cut-off point of or above this would not be the case as proportionally more allegiant studies were conducted in hospital settings.

Similarly, differences in the proportions of studies using translated versions of the PHQ-9 are also unlikely to offer



an obvious explanation of the difference in diagnostic performance, because in the algorithm set of studies more of the allegiant studies used a translated version of the test, but the proportions were in the opposite direction for the studies using a cut-off of 10 or above. We tested this by carrying out a sensitivity analysis restricting the sample to English studies and studies with adequate translation. The allegiance effect was still predictive of DOR variation between allegiance and non-allegiance studies variation in both algorithm ($p=0.00$) and summed item scoring at cut-off point of 10 meta-analyses ($p=0.02$).

A similar conclusion is also likely to apply to the age of the samples. There were more older adults studies in the non-allegiant than allegiant studies in the algorithm comparison. Depression could be more difficult to identify in older adults due to physical comorbidities that may present with similar symptomatology to depression and could account for the lower diagnostic performance in the non-allegiant studies. However, the non-allegiant samples in the studies that reported the psychometric properties at cut-off point 10 or above had younger samples than the allegiant studies, so this would not support this interpretation.

The SCID was used as the gold standard in nearly all allegiant studies. The fact that some non-allegiant studies used other gold standards could potentially explain the poorer psychometric properties of the PHQ-9 in these studies. The SCID is often regarded as the most valid of the available semi-structured interviews used in depression diagnostic validity studies as the reference standard. If we assume that this is the case and, furthermore, that the PHQ-9 is an accurate method of screening for depression, then the PHQ-9 may be more likely to agree with the SCID than other reference standards. However, when we carried out a sensitivity analysis restricting the sample to SCID-only studies, the allegiance effect was still predictive of DOR variation between allegiance and non-allegiance studies variation in both algorithm ($p=0.01$) and summed item scoring at cut-off point of 10 reviews ($p=0.02$).

Differences in methodological quality as potential alternative explanations

The quality of the studies was evaluated using the QUADAS-2. Although there were several potential methodological differences between the two groups of studies from the algorithm papers, not all of these offer obvious explanations of the observed differences and some are unlikely as explanations. For example, more allegiant studies ensured that the reference test was interpreted blind to the index test. This is unlikely to account for the observed differences, because a lack of blinding is typically associated with artificially increased diagnostic performance, which is in the opposite direction to the pattern of results observed here. The impact of some other differences is less clear-cut. For example, a higher number of the non-allegiant studies met the criterion for consecutive referrals. For this to provide an explanation of the observed differences, the non-consecutive nature

of the referrals in the studies by those who had developed the PHQ-9 would need to have led to the overinclusion of true positives or underinclusion of false negatives given that these studies tended to report higher sensitivity relative to the non-allegiant studies (and vice versa for the independent studies). It is not immediately obvious how this would occur. The allegiant studies were more likely to have met the criterion of 'included all participants in the analysis'. It is possible that the greater loss of participants from the non-allegiant studies may have artificially reduced the observed diagnostic accuracy, although, again, it is not immediately obvious how this would have affected the true positive and false negative rates. Although there is not an obvious explanation of how these differences in methodological quality could account for the observed differences in diagnostic performance, it is important to recognise that they cannot on that basis be ruled out.

There are, however, two differences in methodological quality among the algorithm studies that are clearer potential alternative explanations. The higher rate of appropriate translations among the allegiant studies is potentially important, because lower diagnostic estimates may be expected from studies that have poorly translated versions of the index test. In the flow and timing domain, more allegiant studies ensured that there was a less than 2-week interval between the index and reference test. This is consistent with lower diagnostic performance in the non-allegiant studies: as the interval increases it is likely that depression status may change and this would lead to lower levels of agreement between the index test and the reference test.

There were also differences on some quality assessment items between the two sets of studies in the summed item scoring method comparison. The threshold was reported as prespecified in all allegiant studies in contrast to approximately three-quarters of the non-allegiant studies. On the face of it, this is unlikely to explain the observed differences, because the use of a prespecified cut-off point is likely to be associated with lower not higher diagnostic test performance. One possibility, however, is that studies that performed poorly at this cut-off point were less likely to be reported by those who had developed the measure. As discussed in more detail in the 'Limitations' section, we were unable to explore this possibility through the use of formal tests for publication bias.

All allegiant studies avoided inappropriate exclusions compared with approximately half of the non-allegiant studies. While this is a potential alternative explanation of the differences, it is not immediately obvious how this would explain the differences in diagnostic performance between the two sets of studies. Fewer than half of the non-allegiant studies met the criterion for 'all participants included in the analysis', in contrast to all of the allegiant studies met this criterion, but again this difference should usually work against the inclusive studies, not those excluding cases. More of the non-allegiant studies reported that the PHQ-9 was interpreted blind to the



reference test. This does offer a potential explanation, because the absence of blinding may artificially inflate diagnostic accuracy.

LIMITATIONS

The results of this review need to be viewed in light of the limitations of the primary studies that contributed to the review and the review itself. An important consideration is to establish whether any observed differences between the diagnostic performance of the non-allegiant and allegiant studies are better accounted for by study characteristic or methodological differences. Caution, however, is needed in interpreting any differences, because of the small number of allegiant studies in both the algorithm and cut-off 10 or above comparisons. The small number of allegiant studies also meant that we were also unable to explore the potential role of publication bias in the non-allegiant and allegiant studies. At least 10 studies are required to use standard methods of examining publication bias, but the number of allegiant studies in both the algorithm and cut-off 10 or above comparisons were fewer than this. Papers published from August 2013 onwards are not covered in the literature search used and so it potentially misses some more recent studies that would be eligible for inclusion, although it is unlikely that many, if any, new allegiant studies have been published since.

CONCLUSIONS AND IMPLICATIONS FOR FURTHER RESEARCH

The aims of the review was to investigate whether an allegiance effect is found that leads to an increased diagnostic performance in diagnostic validation studies that were conducted by teams connected to the original developers of the PHQ-9. Our analyses showed that diagnostic studies conducted by independent/non-allegiant researchers had lower sensitivity paired with similar specificity compared with studies that were classified as allegiant. This conclusion held for both the algorithm and cut-off 10 or above studies. We explored a range of possible alternative explanations for the observed allegiance effect including both differences in study characteristics and study quality. A number of potential differences were found, although for some of these it is not clear how they would necessarily account for the observed differences. However, there were a number of differences that offered potential alternative explanations unconnected to allegiance effects. In the algorithm studies, the studies rated as allegiant were also more likely to use an appropriate translation of the PHQ-9 and were also more likely to ensure that the index and reference test were conducted within 2 weeks of each other, both of which may be associated with an improvement in observed diagnostic performance of an instrument. The majority of studies in both meta-analyses did not provide clear statements about potential conflict of interest and/or funding; however, the newer

studies were more likely to provide such statements, which may reflect increasing transparency in this area of research.

We cannot, therefore, conclude that allegiance effects are present in studies examining the diagnostic performance of the PHQ-9; but nor can we rule them out. Conflicts of interest are an important area of investigation in medical and behavioural research, particularly due to concerns about trial results being influenced by industry sponsorship. Future diagnostic validity in this area should as a matter of routine present clear statements about potential conflicts of interest and funding, particularly relating to the development of the instrument under evaluation. Future meta-analyses of diagnostic validation studies of psychological measures should routinely evaluate the impact of researcher allegiance in the primary studies examined in the meta-analysis.

Acknowledgements One of the authors of this paper (SG) was supported by the NIHR Collaboration for Leadership in Applied Health Research and Care Yorkshire and Humber (NIHR CLAHRC YH). The views and opinions expressed are those of the author(s), and not necessarily those of the NHS, the NIHR or the Department of Health.

Contributors LM led on all stages of the review and is the guarantor. We used an established database of diagnostic validation studies of the PHQ-9 (Manea *et al.*, 2015; Moriarty *et al.*, 2015). SG provided expert advice on methodology and approaches to assessment of the evidence base. AM carried out the literature searches, screened the studies, extracted data and assessed the quality of the included studies for one of the systematic reviews (Moriarty *et al.*, 2015). LM carried out the literature searches, screened the studies, extracted data and assessed the quality of the included studies for the other systematic review (Manea *et al.*, 2015), analysed the data for both systematic reviews and drafted the report. JB involved in the development of the study, wrote the introduction section of the review and contributed to the production of the final report. DM supervised the quality assessment, methodology and approaches to evidence synthesis, provided senior advice and supported throughout and contributed to the production of the final report. All parties were involved in drafting and/or commenting on the report.

Funding LM was an NIHR Clinical Lecturer when this research was carried out. The NIHR had no role in the study design, methods, data collection, analysis or interpretation of data, nor any role in the preparation of the manuscript or decision to submit the manuscript for publication.

Competing interests None declared.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

Open Access This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: <http://creativecommons.org/licenses/by/4.0/>

© Article author(s) (or their employer(s) unless otherwise stated in the text of the article) 2017. All rights reserved. No commercial use is permitted unless otherwise expressly granted.

REFERENCES

1. Luborsky L, Diguier L, Seligman DA, *et al.* The researcher's own therapy allegiances: a "Wild Card" in comparisons of treatment efficacy. *Clin Psychol: Sci Pract* 2006;6:95–106.
2. Dragioti E, Dimoliatis I, Evangelou E. Disclosure of researcher allegiance in meta-analyses and randomised controlled trials of psychotherapy: a systematic appraisal. *BMJ Open* 2015;5:e007206.

3. Munder T, Brüttsch O, Leonhart R, *et al.* Researcher allegiance in psychotherapy outcome research: an overview of reviews. *Clin Psychol Rev* 2013;33:501–11.
4. Winter DA. "Editorial." *Routledge*, 2010.
5. McLeod J. "Taking allegiance seriously—implications for research policy and practice." *Eur J Psychother Couns* 2010;12.
6. Staines GL, Cleland CM. Bias in meta-analytic estimates of the absolute efficacy of psychotherapy. *Rev. Gen. Psychol* 2007;11:329–47.
7. Markman KD, Hirt ER. Social Prediction and the "Allegiance Bias". *Soc Cogn* 2002;20:58–86.
8. Walters GD. The psychological inventory of criminal thinking styles and psychopathy checklist: screening version as incrementally valid predictors of recidivism. *Law Hum Behav* 2009;33:497–505.
9. Singh JP, Grann M, Fazel S. Authorship bias in violence risk assessment? A systematic review and meta-analysis. *PLoS One* 2013;8:e72484.
10. Blair PR, Marcus DK, Boccaccini MT. Is there an allegiance effect for assessment instruments? actuarial risk assessment as an exemplar. *Clin PsycholSci Pract* 2008;15:346–60.
11. Lilienfeld SO, Jones MK. Allegiance effects in assessment: unresolved questions, potential explanations, and constructive remedies. *Clin PsycholSci Pract* 2008;15:361–5.
12. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606–13.
13. Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the patient health questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *Gen Hosp Psychiatry* 2015;37:67–75.
14. Moriarty AS, Gilbody S, McMillan D, *et al.* Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Gen Hosp Psychiatry* 2015;37:567–76.
15. Whiting PF, Rutjes AW, Westwood ME, *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529–36.
16. Mann R, Hewitt CE, Gilbody SM. Assessing the quality of diagnostic studies using psychometric instruments: applying QUADAS. *Soc Psychiatry Psychiatr Epidemiol* 2009;44:300–7.
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. University of York. NHS Centre for Reviews and Dissemination. *Systematic reviews : CRD's guidance for undertaking reviews in health care*: CRD, University of York, 2009.
19. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002;21:1237–56.
20. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. *Stat Med* 2002;21:1525–37.
21. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21:1559–73.
22. Diez-Quevedo C, Rangil T, Sanchez-Planell L, *et al.* Validation and utility of the patient health questionnaire in diagnosing mental disorders in 1003 general hospital Spanish inpatients. *Psychosom Med* 2001;63:679–86.
23. Gräfe K, Zipfel S, Herzog W, *et al.* Screening psychischer Störungen mit dem "Gesundheitsfragebogen für Patienten (PHQ-D)". *Diagnostica* 2004;50:171–81.
24. Löwe B, Spitzer RL, Gräfe K, *et al.* Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *J Affect Disord* 2004;78:131–40.
25. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA* 1999;282:1737–44.
26. Thekkumpurath P, Walker J, Butcher I, *et al.* Screening for major depression in cancer outpatients: the diagnostic accuracy of the 9-item patient health questionnaire. *Cancer* 2011;117:218–27.
27. Muramatsu K, Miyaoka H, Kamijima K, *et al.* The patient health questionnaire, Japanese version: validity according to the mini-international neuropsychiatric interview-plus. *Psychol Rep* 2007;101(3 Pt 1):952–60.
28. Navinés R, Castellví P, Moreno-España J, *et al.* Depressive and anxiety disorders in chronic hepatitis C patients: reliability and validity of the Patient Health Questionnaire. *J Affect Disord* 2012;138:343–51.
29. Eack SM, Greeno CG, Lee BJ. Limitations of the patient health questionnaire in identifying anxiety and depression: many cases are undetected. *Res Soc Work Pract* 2006;16:625–31.
30. Fann JR, Bombardier CH, Dikmen S, *et al.* Validity of the patient health questionnaire-9 in assessing depression following traumatic brain injury. *J Head Trauma Rehabil* 2005;20:501–11.
31. Gelaye B, Williams MA, Lemma S, *et al.* Validity of the patient health questionnaire-9 for depression screening and diagnosis in East Africa. *Psychiatry Res* 2013;210:653–61.
32. Hyphantis T, Kotsis K, Voulgari PV, *et al.* Diagnostic accuracy, internal consistency, and convergent validity of the Greek version of the patient health questionnaire 9 in diagnosing depression in rheumatologic disorders. *Arthritis Care Res* 2011;63:1313–21.
33. Inagaki M, Ohtsuki T, Yonemoto N, *et al.* 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. *Gen Hosp Psychiatry* 2013;35:592–7.
34. Khamseh ME, Baradaran HR, Javanbakht A, *et al.* Comparison of the CES-D and PHQ-9 depression scales in people with type 2 diabetes in Tehran, Iran. *BMC Psychiatry* 2011;11:61.
35. Persoons P, Luyckx K, Desloovere C, *et al.* Anxiety and mood disorders in otorhinolaryngology outpatients presenting with dizziness: validation of the self-administered PRIME-MD patient health questionnaire and epidemiology. *Gen Hosp Psychiatry* 2003;25:316–23.
36. Picardi A, Adler DA, Abeni D, *et al.* Screening for depressive disorders in patients with skin diseases: a comparison of three screeners. *Acta Derm Venereol* 2005;85:414–9.
37. Stafford L, Berk M, Jackson HJ. 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.
38. 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.
39. Thompson AW, Liu H, Hays RD, *et al.* Diagnostic accuracy and agreement across three depression assessment measures for Parkinson's disease. *Parkinsonism Relat Disord* 2011;17:40–5.
40. Turner A, Hambridge J, White J, *et al.* Depression screening in stroke: a comparison of alternative measures with the structured diagnostic interview for the diagnostic and statistical manual of mental disorders, fourth edition (major depressive episode) as criterion standard. *Stroke* 2012;43:1000–5.
41. van Steenberg-Weijnenburg KM, de Vroeger L, Ploeger RR, *et al.* Validation of the PHQ-9 as a screening instrument for depression in diabetes patients in specialized outpatient clinics. *BMC Health Serv Res* 2010;10:235.
42. Arroll B, Goodyear-Smith F, Crengle S, *et al.* Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med* 2010;8:348–53.
43. Ayalon L, Goldfracht M, Bech P. 'Do you think you suffer from depression?' Reevaluating the use of a single item question for the screening of depression in older primary care patients. *Int J Geriatr Psychiatry* 2010;25:497–502.
44. Henkel V, Mergl R, Kohnen R, *et al.* Use of brief depression screening tools in primary care: consideration of heterogeneity in performance in different patient groups. *Gen Hosp Psychiatry* 2004;26:190–8.
45. Lamers F, Jonkers CC, 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.
46. Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry* 2008;8:46.
47. Zuihoff NP, Vergouwe Y, King M, *et al.* The patient health questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a cross-sectional study. *BMC Fam Pract* 2010;11:98.
48. 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.
49. Williams LS, Brizendine EJ, Plue L, *et al.* Performance of the PHQ-9 as a screening tool for depression after stroke. *Stroke* 2005;36:635–8.
50. Chagas MH, Tumas V, Rodrigues GR, *et al.* Validation and internal consistency of Patient Health Questionnaire-9 for major depression in Parkinson's disease. *Age Ageing* 2013;42:645–9.
51. Elderon L, Smolderen KG, Na B, *et al.* Accuracy and prognostic value of American Heart Association: recommended depression screening in patients with coronary heart disease: data from the Heart and Soul Study. *Circ Cardiovasc Qual Outcomes* 2011;4:533–40.
52. Rooney AG, McNamara S, Mackinnon M, *et al.* Screening for major depressive disorder in adults with cerebral glioma: an initial validation of 3 self-report instruments. *Neuro Oncol* 2013;15:122–9.



53. Watnick S, Wang PL, Demadura T, *et al.* Validation of 2 depression screening tools in dialysis patients. *Am J Kidney Dis* 2005;46:919–24.
54. Zhang Y, Ting R, Lam M, *et al.* Measuring depressive symptoms using the Patient Health Questionnaire-9 in Hong Kong Chinese subjects with type 2 diabetes. *J Affect Disord* 2013;151:660–6.
55. 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.
56. Fine TH, Contractor AA, Tamburrino M, *et al.* Validation of the telephone-administered PHQ-9 against the in-person administered SCID-I major depression module. *J Affect Disord* 2013;150:1001–7.
57. 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.
58. Phelan E, Williams B, Meeker K, *et al.* A study of the diagnostic accuracy of the PHQ-9 in primary care elderly. *BMC Fam Pract* 2010;11:63.
59. Sidebottom AC, Harrison PA, Godecker A, *et al.* Validation of the patient health questionnaire (PHQ)-9 for prenatal depression screening. *Arch Womens Ment Health* 2012;15:367–74.
60. de Lima Osório F, Vilela Mendes A, Crippa JA, *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.
61. 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.
62. Azah MNN, Shah MEM, 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.
63. Liu SI, Yeh ZT, Huang HC, *et al.* Validation of patient health questionnaire for depression screening among primary care patients in Taiwan. *Compr Psychiatry* 2011;52:96–101.
64. 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.

BMJ Open

Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis

Laura Manea, Jan Rasmus Boehnke, Simon Gilbody, Andrew S Moriarty and Dean McMillan

BMJ Open 2017 7:
doi: 10.1136/bmjopen-2016-015247

Updated information and services can be found at:
<http://bmjopen.bmj.com/content/7/9/e015247>

These include:

References

This article cites 59 articles, 7 of which you can access for free at:
<http://bmjopen.bmj.com/content/7/9/e015247#ref-list-1>

Open Access

This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See:
<http://creativecommons.org/licenses/by/4.0/>

Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections

Articles on similar topics can be found in the following collections

[Mental health](#) (791)
[Screening \(epidemiology\)](#) (25)

Notes

To request permissions go to:
<http://group.bmj.com/group/rights-licensing/permissions>

To order reprints go to:
<http://journals.bmj.com/cgi/reprintform>

To subscribe to BMJ go to:
<http://group.bmj.com/subscribe/>