General Hospital Psychiatry 37 (2015) 67–75

Contents lists available at ScienceDirect

General Hospital Psychiatry

journal homepage: http://www.ghpjournal.com

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.

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).

http://dx.doi.org/10.1016/j.genhosppsych.2014.09.009
0163-8343/

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,

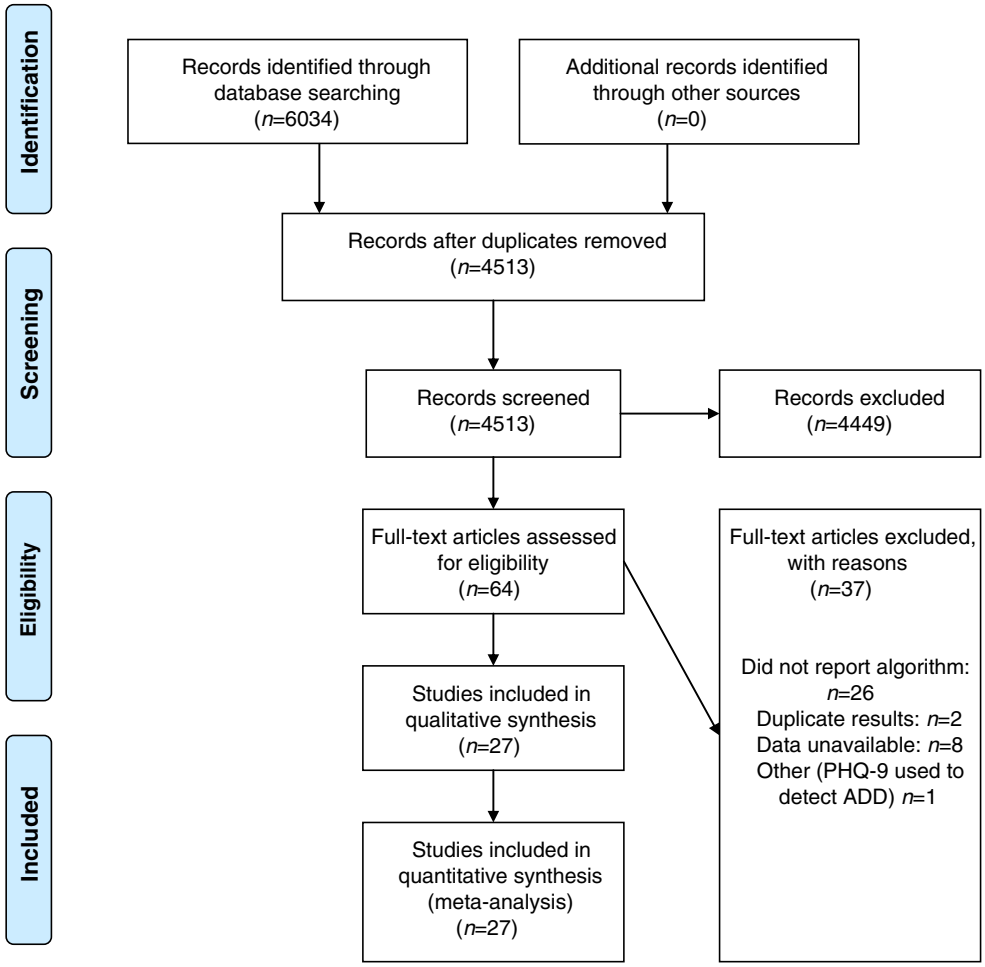

Fig. 1. PRISMA flowchart — search and selection of included diagnostic accuracy studies for systematic review.

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 Steenbergen-Weijenburg 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]	✗	✓	✗	High	?	✓	✓	Unclear
Eack et al. [23]	?	✓	?	Unclear	?	n/a	n/a	Unclear
Fann et al. [24]	✗	✗	✗	High	?	n/a	n/a	Unclear
Gelaye et al. [25]	?	✗	?	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]	✓	✓	✗	High	✓	?	?	Unclear
Inagaki et al. [30]	✓	✗	✓	High	✓	?	?	Unclear
Khamseh et al. [31]	✓	✓	?	Unclear	✓	✓	?	Unclear
Lamers et al. [32]	✓	✗	✗	High	✓	?	?	Unclear
Lotrakul et al. [33]	✗	✓	?	High	✓	✓	?	Unclear
Lowe et al. [34]	✗	✓	✓	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)	✗	✓	✓	High	✓	n/a	n/a	Low
Stafford et al. [39]	✓	✓	✓	Low	✓	n/a	n/a	Low
Thekkumpurath et al. (2010)	✗	✗	✓	High	✓	n/a	n/a	Low
Thombs et al. [41]	✗	✓	?	Unclear	?	n/a	n/a	Unclear
Thomspon et al. (2011)	?	✓	✓	Unclear	?	n/a	n/a	Unclear
Turner et al. [43]	?	✓	✓	Unclear	?	n/a	n/a	Unclear
van Steenbergen-Weijenburg et al. [44]	?	✓	✓	Unclear	?	?	?	Unclear
Zuithoff et al. [45]	✗	✓	✓	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	✓	✓	✗	High
Gelaye et al. [25]	✓	✓	?	?	Unclear	✓	✓	✗	High
Gjerdingen et al. [26]	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Fann et al. [24]	✓	✓	n/a	n/a	Low	✓	✓	✗	High
Gräfe et al. (2004)	✓	?	n/a	n/a	Unclear	✓	✓	✓	Low
Henkel et al. [28]	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Hyphantis et al. [29]	✓	✓	?	?	Unclear	✓	✓	✗	High
Inagaki et el. [30]	✓	✓	✓	?	Unclear	✓	✓	✗	High
Khamseh et al. [31]	✓	✓	✓	?	Unclear	✓	✓	?	Unclear
Lamers et al. [32]	✓	?	?	?	Unclear	?	✓	✗	High
Lotrakul et al. [33]	✓	✓	✓	✓	Low	?	✓	✗	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	✓	✓	✗	High
Spitzer et al. (1999)	✓	✓	n/a	n/a	Low	✓	✓	✗	High
Stafford et al. [39]	✓	✓	n/a	n/a	Low	✓	✓	✗	High
Thekkumpurath et al. (2010)	✓	✓	n/a	n/a	Low	?	✓	✗	High
Thombs et al. [41]	?	✓	n/a	n/a	Unclear	✓	✓	✓	Low
Thompson et al. [42]	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
Turner et al. [43]	✓	?	n/a	n/a	Unclear	✓	✓	✗	High
van Steenbergen-Weijenburg et al. [44]	✓	✗	?	?	High	✓	✓	✗	High
Zuithoff et al. [45]	✓	✓	✓	✓	Low	?	✓	✓	Low

✓, criterion met; ✗, 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 4
Pooled 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

[1] 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.
[2] US Preventive Services Task Force. Guide to Clinical Preventive Services. 2nd edit. Alexandra, VA: International Medical Publishing; 1996.
[3] National Institute for Clinical Excellence. Depression: The treatment and management of depression in adults (updated edition). London: National Institute for Clinical Excellence; 2009.
[4] National Institute for Clinical Excellence. Depression in adults with a chronic physical health problem. London: National Institute for Clinical Excellence; 2009.
[5] Canadian Task Force on Preventive Health Care. Recommendations on screening for depression in adults. Can Med Assoc J 2013;185:775–82.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] Centre for Reviews and Dissemination. Systematic Reviews: CRD's guidance for undertaking reviews in health care. York: University of York; 2009.
[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] Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002;21(11):1559–73.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] 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.
[18] Thompson SG. Systematic review — why sources of heterogeneity in metaanalysis should be investigated. Br Med J 1994;309(6965):1351–5.
[19] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Br Med J 2003;327(7414):557–60.
[20] 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.
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] Henkel V, Mergl R, Kohnen 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.
[29] Hyphantis T, Kotsis K, Voulgari PV, Tsifetaki 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.
[30] 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.
[31] Khamseh 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.
[32] 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.
[33] Lotrakul M, Sumrithe S, Saipanish R. Reliability and validity of the Thai version of the PHQ-9. BMC Psychiatry 2008;8:46.
[34] 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.
[35] 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.
[36] 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.
[37] Persoons P, Luyckx K, Desloovere C, Vandenberghe 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.
[38] 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.
[39] 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.
[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] van Steenbergen-Weijenburg KM, de Vroege L, Ploeger RR, Brals JW, Vloedbeld 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.
[45] 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 crosssectional 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] Wittkampf 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.