Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review


Abstract

Objective
The nine-item mood module of the Patient Health Questionnaire (PHQ-9) was developed to screen and to diagnose patients in primary care with depressive disorders. We systematically reviewed the psychometric literature on the PHQ-9 and performed a meta-analysis to ascertain its summary sensitivity and specificity.

Methods
EMBASE, PubMed and PsycINFO were used to search literature up to July 2006. Studies were included if (1) they investigated the diagnostic accuracy of the PHQ-9 and (2) the PHQ-9 had been compared with a reference test. The quality of the studies was appraised using the Quality Assessment of Diagnostic Accuracy Studies. We calculated sensitivity, specificity and confidence intervals for each included study. We used the random effects model to calculate the summary sensitivity and specificity.

Results
We found a sensitivity of 0.77 (0.71–0.84) and a specificity of 0.94 (0.90–0.97) for the PHQ-9. The positive predictive value in an unselected primary care population was 59%, which increased to 85–90% when the prior probability increased to 30–40%.

Conclusion
In primary care, the PHQ-9 is a valid diagnostic tool if used in selected subgroups of patients with a high prevalence of depressive disorder.
Introduction

Detection of depression, especially in primary care, is far from optimal.\textsuperscript{1-3} Both underdiagnosis and overdiagnosis have been reported, resulting in inadequate treatment. Underdiagnosis is related to the fact that patients present to their general practitioner (GP) with atypical symptoms either because they are too ashamed to discuss psychological problems or because subjective somatic symptoms are the main reason for their consultation. GPs may have difficulty with asking patients frankly about psychological symptoms. Sometimes, they do not know how to introduce the idea that depression may be an explanation for patients' physical complaints.\textsuperscript{4} Overdiagnosis may occur among patients with subclinical depression or in psychological distress who are known to have had earlier episodes of depression.\textsuperscript{5} Underdiagnosis carries the risk that patients do not get effective psychiatric treatment or are inappropriately treated for physical symptoms. Conversely, overdiagnosis carries the risk for unnecessary and therefore ineffective psychiatric treatment of minor and self-limiting problems. An instrument to detect a major depressive episode (MDE) should ideally have both a high sensitivity and a high specificity in order to reduce the number of false-negatives and false-positives.

A number of screening instruments to detect depressive episodes have been developed. Recently, these instruments were evaluated by Williams et al in a literature synthesis reporting similar operating characteristics but differences in administration time, ease of scoring and the ability to serve additional purposes, such as monitoring severity and screening for conditions other than depression.\textsuperscript{6} Of these instruments, only the Patient Health Questionnaire (PHQ) was developed for screening and diagnosis as well as monitoring of depression severity. It was developed in 1999 as a self-report version of the Primary Care Evaluation of Mental Disorders (PRIME-MD) aimed at criteria-based diagnosis not only of depressive episodes but also of other mental disorders commonly encountered in primary care.\textsuperscript{7} Nowadays, the PHQ is used all over the world and has been translated into more than 25 languages, including German, French, Spanish, Italian, Arabic, Bengali, Turkish, Flemish and Dutch.\textsuperscript{8-16}

The nine-item depression module of the full PHQ is called the PHQ-9.\textsuperscript{17} In contrast to other depression questionnaires, the PHQ-9 evaluates the nine Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) criteria for MDE.\textsuperscript{18} The diagnosis of MDE can be made by a categorical algorithm using these nine items. By calculating a summary score, the severity of an episode can be assessed. Several studies have reported on the diagnostic accuracy of this instrument, but so far, the results have not been synthesized. We systematically reviewed the literature on the diagnostic accuracy of the self-report version of the
PHQ-9. We further performed a meta-analysis to calculate its summary sensitivity and specificity.

**Methods**

**Data sources**
We performed a systematic search of literature dating between 1999 (PHQ issued) and July 2006 using the databases EMBASE, PubMed and PsycINFO with the terms “PHQ” and “Patient Health Questionnaire,” both as MESH headings and as text words. In addition, we checked the references of all included articles for relevant studies.

**Study selection**
Articles were included if their titles and abstracts were focused on the diagnostic accuracy of the PHQ-9 for MDE. Furthermore, the PHQ had to have been compared with a reference test based on a structured interview with DSM-IV criteria. There was no language restriction. Articles that met the inclusion criteria were read completely, and the inclusion criteria were rechecked. Studies in which the PHQ-9 was administered by telephone were excluded, as were those that used previously reported data. Two authors (K.W. and L.N.) performed these assessments independently. The final selection was determined in a consensus meeting with a third author (H.W.).

**Quality assessment**
Two reviewers (K.W. and L.N.) appraised each study independently using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS), a validated quality assessment instrument to assess the scientific quality of diagnostic accuracy studies.\(^{19}\) We used the QUADAS because it is short, has already been used in earlier diagnostic reviews and has an established validity.\(^{20-24}\) The QUADAS consists of 14 items on methodological quality that can be scored as “yes”, “no” or “unclear” (Table 1). Scoring was performed in accordance to the QUADAS user's guide.\(^{24}\) In scoring, we scored each of items 1 through 14 as “unclear” if methods were reported incompletely but as “yes” or “no” only if the criterion was explicitly met or not met. Because the PHQ-9 refers to the past 2 weeks and the reference tests refer to the severest 2 weeks in the past 4 weeks, we defined a maximum interval of 2 weeks between the index test and the reference test as acceptable. As scoring of the index test is fully automated and involves no interpretation, we skipped items 10 and 12 in accordance to the user's guide.\(^{24}\) Differences among assessors were resolved at a consensus meeting with one of the authors (H.W.). The limitations of each included study were described.
**Table 1**

The QUADAS tool

1. Was the spectrum of patients representative of the patients who will receive the test in practice?
2. Were selection criteria clearly described?
3. Is the reference standard likely to correctly classify the target condition?
4. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?
5. Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?
6. Did patients receive the same reference standard regardless of the index test result?
7. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?
8. Was the execution of the index test described in sufficient detail to permit replication of the test?
9. Was the execution of the reference standard described in sufficient detail to permit its replication?
10. Were the index test results interpreted without knowledge of the results of the reference standard?
11. Were the reference standard results interpreted without knowledge of the results of the index test?
12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?
13. Were uninterpretable/intermediate test results reported?
14. Were withdrawals from the study explained?

**Data extraction and synthesis**

In performing the meta-analysis, we used the recommendations of other authors. First, a tabulation of study characteristics in which the PHQ version, the reference test, the thresholds to define positive and negative test results and the methodological limitations were shown was presented. The thresholds for positive or negative test results were defined in two ways. Each of the nine questions had four answer categories: “not at all” (zero points); “several days” (one point); “more than half the days” (two points); and “nearly every day” (three points). The diagnostic algorithmic threshold for diagnosing an MDE was regarded as fulfilled if the answer to question 1a and question 1b and five or more of questions 1a–1i was at least “more than half the days” (question 1i was counted if present at all). The threshold PHQ ≥ 10 means that the summary score of questions 1a–1i (range=0–27) has to be 10 or higher. This cut off point was chosen because the
receiver operating characteristic curves in several studies showed this to be the cut off point with the most optimal sensitivity and specificity for screening purposes.\(^{18,31,32}\)

Second, the studies were divided into several subgroups according to sample (primary care or hospital patients) and type of threshold (diagnostic algorithm or PHQ score ≥10). Two-by-two tables were generated from the presented data. If these data could not be generated from the original article, we contacted the authors. From these tables, we calculated the sensitivity, specificity and confidence intervals (CIs) for each study included. Computerized analysis of the data was performed with the use of MetaDisc version 1.2.\(^{33}\) The data were presented in the form of forest plots.

Furthermore, we used the random effects method, a variation on the generic inverse variance method, as described by DerSimonian and Laird, to calculate the summary sensitivity and specificity of the primary care studies.\(^{34}\) The variation on the inverse variance method is to incorporate an assumption that the different studies are estimating different yet related effects. This produces a random effects meta-analysis for which the DerSimonian and Laird method is the simplest version.\(^{35}\) The summary sensitivity and specificity, which are derived from the above analysis, are used to make a prior/post probability figure in which the consequences of a negative or positive PHQ result on the probability of the presence of an MDE are shown.

### Results

#### Study selection

We found 223 articles, of which 40 were selected for detailed reading. Twenty-eight articles were excluded because (1) they did not concern the self-administered version of the PHQ-9 (n=5), (2) the study did not validate the complete mood module of the PHQ (n=4), (3) the PHQ-9 was only used for detecting any depressive episode and not specifically MDE (n=3), (4) the article was not a diagnostic accuracy study on the PHQ-9 (n=14) or (5) the data had been previously published in articles by the same authors (n=2). (For details, see flow diagram 1)

The majority of the included studies (see Table 2 for a description) validated the PHQ-9 in a primary care sample, used the diagnostic algorithm as a threshold to define a positive or negative test result and used the Structured Clinical Interview for DSM-IV Disorders (SCID) as the reference test. Several studies used both thresholds, the diagnostic algorithm and the summary score of the PHQ-9 (mostly
≥10), while some used the summary score only. As expected, the prevalence of MDE varied: it was between 6.3% and 19.8% in the primary care studies, while it varied between 7.9% and 33.3% in the hospital studies.

**Flow diagram 1.** Selection of included diagnostic accuracy studies of the PHQ-9

“Patient Health Questionnaire” OR “PHQ”
N=223

Excluded on title and abstract (N=183):
- No diagnostic accuracy study of the PHQ-9
- Reference test was not based on DSM-IV criteria

Included on title and abstract
N=40

Excluded on full text (N=28):
- No diagnostic accuracy study of the PHQ-9 (N=14)
- No validation of the complete mood module of the PHQ-9 (N=4)
- Another version than the self administered PHQ-9 was used (N=5)
- PHQ-9 was not used to detect major depressive episodes (N=3)
- Data was already presented in former studies (N=2)

Included on full text
N=12
<table>
<thead>
<tr>
<th>StudyID</th>
<th>Population</th>
<th>Prevalence of MDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corapcioglu, 2004&lt;sup&gt;14&lt;/sup&gt;</td>
<td>Adult primary care patients (N=1387)</td>
<td>91/1387=6,6%</td>
</tr>
<tr>
<td>Ez-Quevedo, 2001&lt;sup&gt;10&lt;/sup&gt;</td>
<td>Adult medical and surgical inpatients (N=1003)</td>
<td>83/1003=8,3%</td>
</tr>
<tr>
<td>Grafe, 2004&lt;sup&gt;8&lt;/sup&gt;</td>
<td>Psychosomatic and medical ambulance and primary care adult patients (N=528)</td>
<td>22/357=6,3% 50/171=29,2%</td>
</tr>
<tr>
<td>Kroenke, 2001&lt;sup&gt;17&lt;/sup&gt;</td>
<td>Primary care adult patients (N=580)</td>
<td>41/580=7,1%</td>
</tr>
<tr>
<td>Lowe, 2004&lt;sup&gt;18&lt;/sup&gt;</td>
<td>Primary care adult patients (N=501)</td>
<td>66/501=13,2%</td>
</tr>
<tr>
<td>Mazzotti, 2003&lt;sup&gt;11&lt;/sup&gt;</td>
<td>Dermatological adult inpatients (N=170)</td>
<td>14/170=8,2%</td>
</tr>
<tr>
<td>McManus, 2005&lt;sup&gt;31&lt;/sup&gt;</td>
<td>Coronary Heart disease adult patients (N=1024)</td>
<td>224/1024=21,9%</td>
</tr>
<tr>
<td>Persoons, 2003&lt;sup&gt;15&lt;/sup&gt;</td>
<td>Otorhinolaryngology adult outpatients with dizziness (N=97)</td>
<td>16/97=16,5%</td>
</tr>
<tr>
<td>Picardi, 2005&lt;sup&gt;39&lt;/sup&gt;</td>
<td>Dermatological adult patients (N=141)</td>
<td>11/141=7,8%</td>
</tr>
<tr>
<td>Spitzer, 1999&lt;sup&gt;30&lt;/sup&gt;</td>
<td>Primary care adult patients (N=585)</td>
<td>116/585=19,8%</td>
</tr>
<tr>
<td>Watnick, 2005&lt;sup&gt;32&lt;/sup&gt;</td>
<td>Dialysis adult outpatients (N=62)</td>
<td>16/62=25,8%</td>
</tr>
<tr>
<td>Williams, 2005&lt;sup&gt;36&lt;/sup&gt;</td>
<td>Post stroke adult patients (N=316)</td>
<td>106/316=33,5%</td>
</tr>
</tbody>
</table>

# Structured Clinical Interview for DSM-IV axis I Disorders, *National Institute of Mental Health Diagnostic Interview Schedule, *Mini-International Neuropsychiatric Interview 5.0.0. Key to limitations according to QUADAS [19, 22]: a, Disease progression bias possible, time between index and reference test not described. b, Partial verification bias, part of the sample did not receive the reference test. c, the index test form part of the reference test. d, The execution of the reference test was not described, causing problems with study replication. e, Incorporation bias, the reference test might have been interpreted with knowledge of the index test. f, Unclear if all patients entering the study were accounted for (withdrawals).
<table>
<thead>
<tr>
<th>PHQ</th>
<th>Threshold PHQ</th>
<th>Reference standard</th>
<th>Limitations (QUADAS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete PHQ</td>
<td>Algorithm</td>
<td>Formal interview based on DSM-IV</td>
<td>c, d</td>
</tr>
<tr>
<td>Complete PHQ</td>
<td>Algorithm</td>
<td>Overview of SCID-I</td>
<td>c</td>
</tr>
<tr>
<td>Complete PHQ</td>
<td>Algorithm and summary scores</td>
<td>SCID-I</td>
<td>c</td>
</tr>
<tr>
<td>Mood section</td>
<td>Summary score</td>
<td>SCID-I</td>
<td>c</td>
</tr>
<tr>
<td>Mood section</td>
<td>Algorithm and summary scores</td>
<td>SCID-I</td>
<td>c, f</td>
</tr>
<tr>
<td>Complete PHQ</td>
<td>Algorithm</td>
<td>SCID-I</td>
<td>c, f</td>
</tr>
<tr>
<td>Complete PHQ</td>
<td>Summary score</td>
<td>DIS*</td>
<td>c, e</td>
</tr>
<tr>
<td>Complete PHQ</td>
<td>Algorithm</td>
<td>MINI^</td>
<td>c</td>
</tr>
<tr>
<td>Mood section</td>
<td>Algorithm</td>
<td>SCID-I</td>
<td>c</td>
</tr>
<tr>
<td>Complete PHQ</td>
<td>Algorithm</td>
<td>Overview of SCID-I + PRIME-MD</td>
<td>c</td>
</tr>
<tr>
<td>Mood section</td>
<td>Summary scores</td>
<td>SCID-I Mood module</td>
<td>c</td>
</tr>
<tr>
<td>Mood section</td>
<td>Summary score</td>
<td>SCID-I Mood module</td>
<td>a, b, c, e, f</td>
</tr>
</tbody>
</table>
Quality assessment
The limitations per study in accordance to the quality assessment of the QUADAS are shown in the last column of Table 2. Most of the studies complied with the methodological recommendations of the QUADAS tool. However, several studies revealed methodological shortcomings that could introduce bias, which might have an effect on the results. The first pitfall has to do with the blinding procedure in the interpretation of the tests. McManus et al and Williams et al interpreted their reference test while knowing the results of the PHQ-9.\textsuperscript{31,36} This knowledge probably influenced the outcome of the reference test (review bias).\textsuperscript{19} The reference test's composition as an interview in which the patients' answers are interpreted by the interviewer could result in an overestimation of the various measures of diagnostic accuracy.

A second pitfall is the dependence of the reference test on the index test. In the study by Spitzer et al, the reference test took the form of an interview based on the overview of the SCID with supplementary diagnostic questions taken from the original PRIME-MD.\textsuperscript{30} The PHQ-9 was developed from the original PRIME-MD, and so the reference test is not independent of the index test (incorporation bias). This probably increases the agreement between index test results and the outcome of the reference test, causing diagnostic accuracy to be overestimated. The last pitfall is related to the selection of patients who were given the reference test. In the study by Williams et al, only those patients who scored positive for at least two symptoms or endorsed the depressed mood or anhedonia item on the PHQ-9 were further assessed by the reference test (SCID).\textsuperscript{36} All other patients were considered as not having a depressive episode and were included in the 2×2 table although they were not given the reference test. Using this procedure means the reported specificity will be too high.

Meta-analysis
To perform a meta-analysis, we subdivided the studies according to the threshold, the type of patient sample and the type of reference test. If studies had calculated the sensitivity and specificity for both thresholds, we decided to use only the diagnostic algorithm data to prevent double weighting of samples. Sensitivity, specificity and CIs were calculated for each study. The results are presented in forest plots (figure 2). The sensitivity forest plot shows that the studies of hospital samples in particular are very heterogeneous. This may in part be explained by the differences in patient groups. Seven hospital studies included the following patient groups: post-stroke patients; patients with dizziness; dialysis patients; heart disease patients; dermatology patients; and patients with general medical complaints.\textsuperscript{10,11,15,31,32,36,37} Due to their primary disease, these patients have differing symptoms, some of which may resemble symptoms of depression (e.g., loss of energy, loss of concentration and sleeping problems).
Fig. 2a. Forest plot: Sensitivity (A) and specificity (B) of the included diagnostic accuracy studies on the PHQ-9. Threshold: algorithm.

A

Corapcioglu* (65/91)
Grafe* (13/22)
Lowe* (55/66)
Spitzer* (85/116)
Diez-Quevedo^ (70/83)
Mazzotti^ (6/14)
Persoons^ (11/16)
Picardi^ (6/11)

B

Corapcioglu* (65/91)
Grafe* (13/22)
Lowe* (55/66)
Spitzer* (85/116)
Diez-Quevedo^ (70/83)
Mazzotti^ (6/14)
Persoons^ (11/16)
Picardi^ (6/11)

*Primary care population; ^Secondary care population
Fig. 2b. Forest plot: Sensitivity (C) and specificity (D) of the included diagnostic accuracy studies on the PHQ-9. Threshold: PHQ-9 summary score ≥ 10.

*Primary care population; ^Secondary care population
The sensitivity figures of the studies by McManus et al, Mazzotti et al and Picardi et al in particular differed from those of the other studies.\textsuperscript{11,31,37} We cannot fully explain the variability within the results found by McManus et al.\textsuperscript{31} This relatively large study showed a pitfall in the methodology it used, however: the interviewers were not blinded to the test results. This bias would logically increase the similarity between the index test and the reference test and therefore increase the sensitivity. The authors of the other studies explained that the variability in the Italian version of the PHQ-9 was due to differences caused by translation and to transcultural adaptation problems.\textsuperscript{11,37} Another explanation of this heterogeneity could be related to differences in administration of reference test (SCID), differing disease spectra and samples (inpatients vs. outpatients), differences in test methods or reference test and study quality.

The random effects model was used to pool the studies with the lowest heterogeneity — the primary care studies that used the algorithmic threshold. Within this subgroup, only the study by Spitzer et al showed a remarkably high specificity, which differed significantly from the other specificity values.\textsuperscript{30} Therefore, we decided to pool with and without the results found by Spitzer et al to incorporate this heterogeneity into our analysis. Table 3 shows the results of this random effects model. Summary sensitivity and specificity were 0.77 (0.71–0.84) and 0.94 (0.90–0.97), respectively, in all four studies as well as 0.80 (0.70–0.89) and 0.92 (0.90–0.94), respectively, in all studies excluding the study by Spitzer et al. As these data show that the influence of the study by Spitzer et al is very small, we decided to include it in our meta-analysis.

<table>
<thead>
<tr>
<th>Study</th>
<th>Sens</th>
<th>95% CI</th>
<th>TP/(TP+FN)</th>
<th>Spec</th>
<th>95% CI</th>
<th>TN/(TN+FP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corapcioglu, 2004</td>
<td>0.71</td>
<td>0.62–0.81</td>
<td>65/91</td>
<td>0.92</td>
<td>0.90–0.93</td>
<td>1191/1296</td>
</tr>
<tr>
<td>Grafe, 2004</td>
<td>0.86</td>
<td>0.72–1.01</td>
<td>19/22</td>
<td>0.94</td>
<td>0.91–0.97</td>
<td>315/335</td>
</tr>
<tr>
<td>Lowe, 2004</td>
<td>0.83</td>
<td>0.74–0.92</td>
<td>55/66</td>
<td>0.90</td>
<td>0.87–0.83</td>
<td>392/435</td>
</tr>
<tr>
<td>Spitzer, 1999</td>
<td>0.73</td>
<td>0.65–0.81</td>
<td>85/116</td>
<td>0.98</td>
<td>0.97–0.99</td>
<td>459/469</td>
</tr>
<tr>
<td>Pooled</td>
<td>0.77</td>
<td>0.71–0.84</td>
<td></td>
<td>0.94</td>
<td>0.90–0.97</td>
<td></td>
</tr>
<tr>
<td>Heterogeneity chi-squared (Q) =5.75 (p=0.219)</td>
<td>Heterogeneity chi-squared (Q) =46.72 (p=0.000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Pooled*       | 0.80  | 0.70–0.89  |            | 0.92  | 0.90–0.94  |            |
| Heterogeneity chi-squared (Q) =4.46 (p=0.107) | Heterogeneity chi-squared (Q) =4.21 (p=0.122) |

*Analysis without Spitzer because of significant heterogeneity in specificity

TP=True-Positives. FN=False Negatives. TN=True Negatives. FP=False Positives
Fig. 3 shows the prior/post probability of MDE with the use of the PHQ-9 (algorithmic threshold). In drawing the figure, we calculated positive and negative predictive values for different prior probabilities with the use of the summary sensitivity and specificity of the random effects model of the four studies. This figure can easily be used in clinical practice to estimate the probability of the presence of MDE in cases with a positive PHQ-9 result. For instance, if a patient has a prior PHQ-9 probability of MDE of 10% and shows a positive result on the PHQ-9, the probability of MDE increases to 59%. If this patient has a negative result on the PHQ-9, the probability drops from 10% to 3%. If a patient has a prior probability of MDE of 30–40%, the post-test probability increases to 85–90% at the cost of missing an MDE with a probability of about 10%.
Discussion

Our meta-analysis, in which we used the random effects model to calculate the summary sensitivity and specificity of four primary care studies, shows that the PHQ-9 has a high specificity of 0.94 (range=0.90–0.97) when used with the algorithm. This indicates that the PHQ-9 is a reliable tool if the user wants to avoid overdiagnosis. On the other hand, the chance of missing a patient with a depressive disorder in an unselected primary care sample (estimated prevalence of MDE=10%) is substantial because the overall sensitivity of 0.77 (range=0.71–0.84) is quite low. If the user wants to avoid missing patients with an MDE, a cut-off point of ≥10 might be more useful than the diagnostic algorithm (sensitivity=0.88, specificity=0.88), although this is based on only one primary care study. Because of the lower specificity, the risk for overdiagnosis however increases and further diagnostic processes are warranted. Overdiagnosis is less likely among patients with a high prior probability. The performance of the PHQ-9 is therefore better in patient samples with a relatively high prevalence of a major depressive disorder. The sensitivity of the PHQ-9 is similar to the characteristics of other case-finding instruments; its specificity, however, is slightly better. Williams et al reported an overall sensitivity of 0.79 (CI=0.74–0.83) and an overall specificity of 0.75 (CI=0.70–0.81) in their literature synthesis.

It is interesting that reported specificity values are much more similar than are the sensitivity values. One explanation for this phenomenon is that the prevalence of depression in primary care samples is relatively low (median=9.7%), which results in wider CIs for the sensitivity figures than for the specificity figures. Another argument could be that the symptoms of depressive patients better meet the DSM-IV criteria if an interviewer adapts and/or explains these criteria to them. Because an interview like the SCID allows for more opportunities to clinically judge a symptom as compared with the “one phrase for all” method of the PHQ-9, this might result in an increased number of false-negatives on the PHQ-9.

In comparing the PHQ-9 with the reference standard, the DSM-IV criteria, we observed one problem in the studies that had previously gone unnoticed. The PHQ-9 detects patients with a depressive episode, while the SCID detects patients with a depressive disorder. The DSM-IV exclusion criteria for a depressive disorder are not included in the PHQ-9. For instance, if a depressive episode is part of a bipolar disorder, the patient has a bipolar disorder and not a depressive disorder according to the SCID. The PHQ-9, on the other hand, does not use a hierarchical structure and will therefore detect an MDE regardless of the presence of a bipolar disorder. Comparing the PHQ-9 with the SCID (and therefore comparing depressive episodes with depressive disorders) will result in false-positives and will therefore lower the specificity of the PHQ-9. This difference between the PHQ-9
and the reference test hampers the use of the PHQ-9 as a diagnostic instrument and should be resolved by further diagnostic workup, most likely by a professional.

The QUADAS tool, which we used to assess the quality of the studies, is not fully applicable to diagnostic studies on a syndrome diagnosis like depressive disorder. This is particularly relevant to item 7, the requirement that the index test should not form part of the reference test. In fact, every diagnostic instrument focuses on the same depressive symptoms mentioned in the DSM-IV. Both index and reference tests are derived from the same diagnostic classification and hence are not fully independent of each other. Therefore, item 7 is not applicable to studies that validate diagnostic tools, as is the case in most psychiatric diagnoses.

The limitations of our study are fourfold. First, because of the large heterogeneity, we only pooled the results of four primary care studies. The outcomes of this meta-analysis are therefore only applicable to primary care samples. Second, the random effects model to pool the sensitivity and specificity values was not able to weigh the relation between the sensitivity and specificity that exists within each study. Reitsma et al developed the bivariate meta-regression model to incorporate this relationship in calculating the summary sensitivity and specificity. We tried to perform this analysis; however, because there were too few studies, the program was not able to calculate the relationship between sensitivity and specificity. We recommend the use of a bivariate meta-regression model in future diagnostic accuracy meta-analyses that include more studies to calculate more specific summary sensitivity and specificity values. Third, as with any review, we cannot rule out publication bias. We might have missed some studies despite our sensitive search, and unpublished studies were not included in this meta-analysis. In addition, our quality assessment does not guarantee bias-free results. Bias may have been introduced into the results of the primary studies and should be kept in mind when interpreting the results of this review.

We conclude that our meta-analysis shows that the PHQ-9 is a valid instrument to detect patients with an MDE. However, in samples with a low pre-test probability of depression (10%), this instrument, like any other depression measure, is not suitable for diagnosing MDE, with the risk for overdiagnosis and overtreatment of MDE. If the PHQ-9 is used in these samples, this instrument, which only suggests “possible” MDE in patients with a positive test score, should be followed by further diagnostic processes. These diagnostic processes should include a diagnostic interview by which somatic causes of the depressive symptoms can be ruled out. In addition, one other DSM-IV criterion for the diagnosis of a depressive disorder, namely of the symptoms causing clinically significant distress or impairment in social, occupational or other important areas of functioning, should in some way be included and explored when making a diagnosis of depressive disorder. On the
other hand, if the PHQ-9 is used in subgroups of patients with an increased risk for depression (prevalence=30–40%), the risk of over-diagnosing is acceptable. These subgroups of patients could be patients who frequently consult their GP or those who have medically unexplained somatic complaints. These subgroups are known to be at an increased risk of having or developing psychiatric disorders.\textsuperscript{38,39} The PHQ-9 can therefore contribute to the improvement of diagnosing depressive episodes in specified primary care samples.

**Acknowledgments**

The authors thank Rob J.P.M. Scholten, MD, PhD, for his statistical advice and for constructing Figure 3 of this article.
References


