Major depressive disorder in primary care: screening, diagnosis and treatment
Baas, K.D.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
CHAPTER 3

The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care


Abstract

Objective Only half of patients with depressive disorder are diagnosed by their general practitioner. Screening in high-risk groups might reduce this hidden morbidity. This study aims to determine the accuracy of the Patient Health Questionnaire-9 (PHQ-9) in (a) screening for depressive disorder, (b) diagnosing depressive disorder and (c) measuring the severity of depressive disorder in groups that are at high-risk for depressive disorder.

Method We compared the performance of the PHQ-9 as a screening instrument and as a diagnostic instrument to that of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) interview, which we used as reference standard. Three high-risk groups of patients were selected: (a) frequent attenders, (b) patients with mental health problems and (c) patients with unexplained somatic complaints. Patients completed the PHQ-9. Next, patients who were at risk for depression (based on PHQ scores) and a random sample of 20% of patients who were not at risk were selected for a second PHQ-9 and the reference standard (SCID-I). We assessed the adequacy of the PHQ-9 as a tool for severity measurement by comparing PHQ-9 scores with scores on the 17-item Hamilton Depression Rating Scale (HDRS-17) in patients diagnosed with a depressive disorder.

Results Among 440 patients, both PHQ-9 and SCID-I were analyzed. The test characteristics for screening were sensitivity=0.93 and specificity=0.85; those for diagnosing were sensitivity=0.68 and specificity=0.95. The positive likelihood ratio for diagnosing was 14.2. The HDRS-17 was administered in 49 patients with depressive disorder. The Pearson correlation coefficient of the PHQ-9 to the HDRS-17 was r=.52 (P<.01).

Conclusion The PHQ-9 performs well as a screening instrument, but in diagnosing depressive disorder, a formal diagnostic process following the PHQ-9 remains imperative. The PHQ-9 does not seem adequate for measuring severity.
Validity of the PHQ-9

Introduction

Improvement in the detection of depressive disorder in primary care is an important objective, since population-based surveys revealed that, currently, only about half of patients with depressive disorder are detected by regular health care providers. Screening has been proposed as a possible solution to this problem. However, the positive predictive value of available screening instruments is too low in unselected groups of patients. One explanation for this is the relatively low prevalence of undisclosed depressive disorder in these unselected populations. Hence, screening yields many false-positive results. Screening in high-risk groups seems to be a reasonable alternative in reducing this number of false positives.

Instruments that are used to screen for depressive disorder are numerous and, overall, have similar and acceptable psychometric properties. The Patient Health Questionnaire-9 (PHQ-9), however, was developed not only for screening for depressive disorder but also for diagnosing depressive disorder. The PHQ-9 is the mood module of the PHQ and consists of nine items. Specific items on the PHQ are derived from The Primary Care Evaluation of Mental Disorders (PRIME-MD) interview schedules and are designed to establish psychiatric diagnoses based on Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria.

Diagnosing depressive disorder with the PHQ-9 is possible because all nine DSM-IV criteria of a depressive episode are included. These screening and diagnostic features have not been validated before within a high-risk population in primary care. A third characteristic of the instrument is the assessment of the severity of a diagnosed depressive disorder, which might help to guide treatment decisions and allows monitoring of treatment effects. The latter has been recommended not only for mental health care settings but also for primary care. The performance of the PHQ-9 in measuring severity has not been well established.

Furthermore, it is unclear whether co-morbid psychiatric disorders such as anxiety and somatoform disorders influence the accuracy of the PHQ-9 in diagnosing depressive disorder. Many patients suffer from both anxiety and depressive symptoms. Besides, some depressive symptoms are similar to anxiety and somatoform symptoms. Therefore, patients with co-morbid psychiatric symptoms might have higher scores on the PHQ-9 compared to patients without co-morbid symptoms, thereby decreasing specificity. This phenomenon has not been studied before.
The PHQ-9 is being used extensively in screening for depressive disorder, diagnosing depressive disorder and assessing depressive disorder severity in primary health care.\textsuperscript{9} Yet evidence to support its use in everyday practice is limited. In this study, we aim to establish the accuracy of the PHQ-9 for four different purposes. Firstly, we aim to validate the screening mode of this instrument within a high-risk population for depressive disorder in primary care, as screening in unselected populations is considered ineffective.\textsuperscript{10} Secondly, we will study the diagnostic accuracy of the instrument by using the PHQ-9 algorithm because diagnosing depressive disorder with an easy-to-use standardized instrument such as the PHQ-9 could support the diagnostic process of general practitioners (GPs). Thirdly, we will study the influence of co-morbid anxiety and somatoform disorder on the test characteristics of the PHQ-9 by performing a subgroup analysis for patients who scored positive on the PHQ-anxiety module or somatoform disorder module. Fourthly, we will assess the ability of the PHQ-9 to measure the severity of depressive disorder.

**Methods**

We studied the screening and diagnostic test characteristics of the PHQ-9 within a high-risk population for depressive disorder in primary care. We compared the results of the PHQ-9 with the results of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I), which was used as reference standard. The SCID-I is a diagnostic interview for DSM-IV Axis I diagnoses.\textsuperscript{11} To study the ability to measure severity, we compared the PHQ-9 with the 17-item Hamilton Depression Rating Scale (HDRS-17).\textsuperscript{12-14}

**Setting**

The study was conducted from March 2006 through July 2007 in primary care patients aged between 18 and 70 years old, in two regions connected to academic settings in The Netherlands: the Academic Medical Center in Amsterdam and the Radboud University Nijmegen Medical Center. The institutional ethics review committee of both centers approved the study protocol.

**Patients**

In The Netherlands, virtually every inhabitant is registered with a GP, who is consulted in case of medical (somatic and psychiatric) needs and acts as a gatekeeper for specialist care. Three groups of patients were selected from a list of 23 GPs. Specific details of the selection of these three high-risk groups can be found in Baas et al.:\textsuperscript{15} 1. Patients with unexplained somatic complaints (USC group). According to their GP, these complaints cannot be explained in medical terms. These complaints had to be
present for at least 3 months. GPs selected these patients from consultation lists within 4 weeks.

2. Patients with consultation rates in the highest 10% in the year preceding study allocation (frequent attenders (FA) group). We used the method proposed by Howe et al.\textsuperscript{16} We determined the highest 10% consulting patients separately for each GP to account for differences in practice styles.

3. Patients presenting to their GP with a new mental health problem up to 3 months prior to the selection date (mental health problem (MHP) group). We selected these patients from electronic patient data-bases of participating GP practices. Three months was chosen as the timeframe because of the transitory nature of most mental health problems.

**Procedures**

GPs received a list of selected patients and were asked to exclude those patients who were suffering from schizophrenia, psychosis or bipolar disorder, as these conditions prevail over depressive disorder according to DSM-IV. Moreover, GPs excluded patients who were considered by the GP as not qualified to enter the study because of specific somatic problems (e.g., blindness, terminal illness) or mental retardation. A third reason for exclusion was trouble with the Dutch or English language. Subsequently, an invitation letter describing the study, signed by the patients’ own GP, was sent to the patients. The letter also contained an informed consent form and the PHQ. If patients did not respond within 2 weeks, a reminder was sent. Missing values on the PHQ were solved by telephone contact.

We used a two-stage design because of efficiency reasons.\textsuperscript{17} Previous PHQ-9 accuracy studies have presented wide confidence intervals for the sensitivity of the PHQ-9 due to the low prevalence of cases, even after meta-analysis.\textsuperscript{3,18} These studies have already reported that patients with a PHQ-9 sum score below 5 are at very low risk for depressive disorder (negative predictive value: 0.99). Therefore, we decided to include a random sample of these low-risk patients for application of the reference standard (the SCID-I interview).

All participants completed the complete PHQ screening questionnaire and were divided into three categories based on PHQ results: (a) PHQ positives (positive score on PHQ-mood module, PHQ-panic syndrome, PHQ-other anxiety or PHQ-somatization disorder); (b) PHQ-9 sum score of 5 or higher (and no other PHQ disorder); and (c) PHQ-9 sum score below 5 and no other PHQ disorder.\textsuperscript{5} We invited all patients from the first two categories and a random sample of 20% of patients from the third category for a second complete PHQ and SCID-I interview. The second PHQ was sent to patients who
had been reached for an appointment for the SCID-I. Patients were asked to fill in the second PHQ just before the SCID-I interview. This second PHQ was used for the analysis. The ability of the PHQ-9 to ascertain the severity of a depressive disorder was assessed by comparison with the HDRS-17. This was performed in patients suffering from depressive disorder, according to the SCID-I, who gave informed consent for an additional HDRS-17 measurement.

**Measurement instruments**

*PHQ and PHQ-9*

The PHQ was developed to make a criteria-based diagnosis not only of a depressive disorder but also of other mental disorders commonly encountered in primary care (anxiety disorders, somatoform disorders, eating disorders and substance abuse). In this study, we only used the depression, anxiety and somatoform disorder modules. The nine-item depression module (PHQ-9) is one part of the PHQ. In contrast to other depression questionnaires, each item of the PHQ-9 evaluates the presence of one of the nine DSM-IV criteria of a depressive episode in the past 2 weeks: (a) depressed mood, (b) anhedonia, (c) trouble sleeping, (d) feeling tired, (e) change in appetite or weight, (f) guilt or worthlessness, (g) trouble concentrating, (h) feeling slowed down or restless, (i) suicidal thoughts. There are four answer categories: 0 (not at all), 1 (few days), 2 (more than half the days) and 3 (almost every day).

In addition to the sum score (0–27) that is used for screening purposes and for measuring depression severity, the PHQ-9 offers a categorical algorithm for the diagnosis of depressive disorder, based on the modified criteria of depressive disorder according to the DSM-IV. A positive outcome on the categorical algorithm requires that five or more of the nine depressive symptom criteria are present more than half the days in the past 2 weeks (suicidal thoughts count if present at all), and one of these symptoms has to be depressed mood or anhedonia.

*SCID-I*

The SCID-I is a semi-structured interview used for diagnosing mental disorders on Axis I according to DSM-IV criteria. The SCID-I was administered by researchers after they had been trained by a skilled professional. Throughout the study, all interviewers had ongoing training sessions supervised by an expert psychiatrist. They also had monthly consensus meetings to maximize accuracy and consistency in the administration of the SCID-I. This SCID-I interview was administered by telephone. The interviewers were blinded to the PHQ-9 results. The agreement between diagnosis gained from telephone contact and live administration of the SCID-I has been found to be excellent ($\kappa=0.73$, with 90% agreement).
HDRS-17
Measurement of depression severity was performed with the HDRS-17, which is a widely used interview-based instrument. The HDRS-17 can be considered as the reference standard used to assess the severity of depressive disorder. We used the 17-item version with a maximum sum score of 52 points. A sum score of 0–7 represents minimal depressive episode; a score of 8–13 represents a mild depressive episode; a score of 14–18 represents a moderate depressive episode; and a score of 19–52 represents a severe depressive episode (>23: a very severe depressive episode). This interview was administered by telephone contact after the SCID-I interview. The agreement between diagnosis gained from telephone contact and live administration of the HDRS-17 has been found to be excellent (intraclass correlation=.80).

Statistical analysis
Prevalence
For the calculation of the prevalence of depressive disorder, we had to correct for the fact that we did not apply the reference standard to all PHQ-9-negative patients. We corrected this imbalance with inverse probability weighting (IPW) and multiplied all patients from the random sample with a factor of 5.21 All calculations have been performed with these corrected data because, otherwise, the disease prevalence and test characteristics (sensitivity and specificity) will be different as a result of our sampling procedure.

Criterion validity
We assessed the criterion validity of the second PHQ-9 by calculating the test characteristics (sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio (LR) and negative LR) using different cutoff values. For the assessment of screening properties, we used the sum scores of the PHQ-9; for the assessment of diagnostic properties, we used the diagnostic cutoff value. Ninety-five percent confidence intervals for proportions were calculated according to the efficient score method (corrected for continuity).22,23 We visualized summary values in a receiver operating characteristic (ROC) curve. Patients with missing values (n=5) on the PHQ-9 were excluded from an analysis of the validity of the PHQ-9 as a screening instrument because no sum score could be calculated. These five patients could be included in the calculation of algorithmic outcome because only one or two values were missing, and these missing values did not change the outcome of the categorical algorithm (PHQ-9 positive or PHQ-9 negative). We performed a subgroup analysis among patients with a positive score on PHQ-anxiety disorder (panic disorder and/or other anxiety disorder) or somatoform disorder to find out whether the test characteristics of the PHQ-9 differed for patients with co-morbid symptoms.
Severity measurement
We explored the relation of PHQ-9 scores and HDRS-17 scores with scatter plot and regression line. Secondly, we calculated the Pearson correlation coefficient to calculate the relationship between PHQ-9 sum score and HDRS-17 score. We consider a Pearson correlation coefficient of .7 as good agreement.

Results
Study population
Six health centers with 23 GPs and 31915 enlisted patients participated. In total, 2659 patients (8.3%) fulfilled the criteria for frequent attending (FA; 1745), mental health problems (MHP; 1049) and/or unexplained somatic complaints (USC; 183). GPs excluded 345 (13.0%) patients. Among the remaining 2314 patients (FA: 1474 (63.7%); MHP: 940 (40.6%); USC: 168 (7.3%)), 1029 (44.5%) patients returned the questionnaire, of which 980 (42.4%) patients gave informed consent. Consenting patients (n=980) were older (mean age, 49.7 vs. 44.2 years, P<.05), were more often female (females: 64% vs. 59%, P<.05) and belonged less often to the MHP group, compared to non-responders (38.2% vs. 42.5%, P<.05). Considering the other two high-risk groups (FA: 65.6% vs. 62.3%; USC: 7.1% vs. 7.4%, P≥.05), we found no statistically significant differences.

We were able to contact 479 (73%) of the 659 patients who qualified for the second PHQ and SCID-I. Valid scores were obtained from 440 patients (FA: 287 (65.2%); MHP: 179 (40.7%); UCS: 33 (7.5%)). A considerable number of patients belonged to more than one high-risk group (see Figure 1). Seventy percent was female, and the mean age was 49.2 years (see Table 1). Comparing patients who were included for the final analysis (n=440) with patients who could not be reached or had incomplete PHQ data (n=216), we found that only mean age differed significantly (49.2 vs. 46.7 years, P=.01). We found no significant differences in terms of sex (69.5% vs. 68.5% females, P≥.05); distribution among high-risk groups (FA: 65.2% vs. 65.3%; MHP: 40.7% vs. 35.6%; USC: 7.5% vs. 11.4%; all P≥.05); and mean PHQ-9 sum score (9.6 vs. 9.5, P≥.05). Fifty-six of 440 patients initially had a PHQ-9 score of <5; for the final analysis, these 56 patients were weighted with a factor of 5 (56×5=280) because of the sampling procedure within this category.21 All following calculations were performed with the corrected population (n=280+384=664) (see flowchart in Figure 2).

Diagnoses according to SCID-I
The prevalence of depressive disorder in our high-risk population (calculated with IPW) was 12.3% (82 of 664 patients), with a 95% confidence interval (95% CI) of 9.9–15.2.
Validity of the PHQ-9

### Table 1. Study population (n=664)

<table>
<thead>
<tr>
<th>Population</th>
<th>Patients (n(%))</th>
<th>Female (n(%))</th>
<th>Mean age (Years)</th>
<th>Depressive Disorder (n(%))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study population</td>
<td>664 (100.0%)</td>
<td>443 (66.7%)</td>
<td>49.8</td>
<td>82 (12.3%)</td>
</tr>
<tr>
<td>- FA&lt;sup&gt;a&lt;/sup&gt;</td>
<td>455 (68.5%)</td>
<td>280 (61.5%)</td>
<td>49.9</td>
<td>58 (12.7%)</td>
</tr>
<tr>
<td>- MHP&lt;sup&gt;a&lt;/sup&gt;</td>
<td>239 (36.0%)</td>
<td>171 (71.5%)</td>
<td>47.5</td>
<td>33 (13.8%)</td>
</tr>
<tr>
<td>- USC&lt;sup&gt;a&lt;/sup&gt;</td>
<td>41 (6.2%)</td>
<td>37 (90.2%)</td>
<td>50.9</td>
<td>8 (19.5%)</td>
</tr>
<tr>
<td>- FA and MHP&lt;sup&gt;a&lt;/sup&gt;</td>
<td>54 (8.1%)</td>
<td>30 (55.6%)</td>
<td>44.1</td>
<td>14 (25.9%)</td>
</tr>
<tr>
<td>- FA and USC&lt;sup&gt;a&lt;/sup&gt;</td>
<td>16 (2.4%)</td>
<td>14 (87.5%)</td>
<td>54.5</td>
<td>3 (18.8%)</td>
</tr>
<tr>
<td>- MHP and USC&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3 (0.5%)</td>
<td>3 (100%)</td>
<td>50.7</td>
<td>1 (33.3%)</td>
</tr>
<tr>
<td>- FA and MHP and USC</td>
<td>2 (0.3%)</td>
<td>2 (100%)</td>
<td>55.5</td>
<td>1 (50%)</td>
</tr>
</tbody>
</table>

<sup>a</sup>A considerable number of patients belonged to more than one high-risk group

Patients belonging to the USC group had the highest prevalence of depressive disorder (19.5%). When a patient belonged to more than one high-risk group, the prevalence increased to 18.8% (FA and USC), 25.9% (FA and MHP) and 33.3% (MHP and USC) (see also Table 1 and Figure 1).
Chapter 3

Figure 2. Flowchart of study population

*IPW: Inversed probability weighting
Test characteristics for screening purposes
ROC curve analysis indicated that the PHQ-9 performed well: PHQ-9 area under the curve=0.95 (95% CI: 0.93–0.97) (see ROC curve; Fig. 3). The cutoff point that is most widely used to indicate a positive case for depressive disorder is the sum score of 10 or higher. Of the 81 SCID-I depressed patients, 75 patients had a total PHQ-9 score of 10 or higher. Of the 578 SCID-I non-depressed patients, 492 patients scored negative on this cutoff point (five missing values). This resulted in a PHQ-9 sensitivity of 0.93 (95% CI: 0.84–0.97), a specificity of 0.85 (95% CI: 0.82–0.88), a positive predictive value of 0.47, a negative predictive value of 0.99, a positive LR of 6.2 (95% CI: 5.1–7.6) and a negative LR of 0.09 (95% CI: 0.04–0.2). For details about other cutoff points, see Table 2A.

Test characteristics for diagnostic purposes
Of the 82 SCID-I depressed patients, 56 patients scored positive on the PHQ-9 according to the diagnostic algorithm. Of the 582 SCID-I non-depressed patients, 554 patients scored negative on this cutoff point. This resulted in a PHQ-9 sensitivity of 0.68 (95% CI: 0.57–0.78), a specificity of 0.95 (95% CI: 0.93–0.97), a positive predictive value of 0.67, a negative predictive value of 0.96, a positive LR of 14.2 (95% CI: 9.6–21.0) and a negative LR of 0.3 (95% CI: 0.2–0.5) (see also Table 2A).

Figure 3. ROC-curve of PHQ-9 sum scores.
### Table 2a. Test characteristics of PHQ-9 corrected with inversed probability weighting.

<table>
<thead>
<tr>
<th>Cutoff point</th>
<th>total</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>sensitivity</th>
<th>specificity</th>
<th>ppv</th>
<th>npv</th>
<th>LR+</th>
<th>LR-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sample corrected with inversed probability weighting</td>
<td></td>
</tr>
<tr>
<td>Algorithm</td>
<td>664</td>
<td>56</td>
<td>28</td>
<td>26</td>
<td>554</td>
<td>.68(.57-.78)</td>
<td>.95(.93-.97)</td>
<td>.67</td>
<td>.96</td>
<td>14.2(9.6-21.0)</td>
<td>0.3 (0.2-0.5)</td>
</tr>
<tr>
<td>PHQ≥10</td>
<td>659*</td>
<td>75</td>
<td>86</td>
<td>6</td>
<td>492</td>
<td>.93(.84-.97)</td>
<td>.85(.82-.88)</td>
<td>.47</td>
<td>.99</td>
<td>6.2(5.1-7.6)</td>
<td>0.1 (0.04-0.2)</td>
</tr>
<tr>
<td>PHQ≥15</td>
<td>659*</td>
<td>53</td>
<td>24</td>
<td>28</td>
<td>554</td>
<td>.65(.54-.75)</td>
<td>.96(.94-.97)</td>
<td>.69</td>
<td>.95</td>
<td>15.8(10.3-24.0)</td>
<td>0.4 (0.3-0.5)</td>
</tr>
</tbody>
</table>

a: true positives; b: false positives; c: false negatives; d: true negatives; ppv, positive predictive value; npv, negative predictive value; LR, likelihood ratio.

*Missings (n=5) of which the sum score could not be calculated.

### Table 2b. Test characteristics of PHQ-9 of patients with co-morbid anxiety disorder or somatoform disorder.

<table>
<thead>
<tr>
<th>Cutoff point</th>
<th>total</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>sensitivity</th>
<th>specificity</th>
<th>ppv</th>
<th>npv</th>
<th>LR+</th>
<th>LR-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sample corrected with inversed probability weighting</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Co-morbid anxiety disorder</td>
<td></td>
</tr>
<tr>
<td>Algorithm</td>
<td>93</td>
<td>39</td>
<td>8</td>
<td>7</td>
<td>39</td>
<td>.85(.71-.93)</td>
<td>.83(.69-.92)</td>
<td>.83</td>
<td>.85</td>
<td>5.0(2.6-9.5)</td>
<td>0.2 (0.09-0.4)</td>
</tr>
<tr>
<td>PHQ≥10</td>
<td>92*</td>
<td>45</td>
<td>26</td>
<td>0</td>
<td>21</td>
<td>1(.90-1)</td>
<td>.45(.30-.60)</td>
<td>.63</td>
<td>1</td>
<td>1.8(1.4-2.3)</td>
<td>0 (np)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Co-morbid somatoform disorder</td>
<td></td>
</tr>
<tr>
<td>Algorithm</td>
<td>137</td>
<td>42</td>
<td>14</td>
<td>4</td>
<td>77</td>
<td>.91(.78-.97)</td>
<td>.85(.75-.91)</td>
<td>.75</td>
<td>.95</td>
<td>5.9(3.6-9.7)</td>
<td>0.1(0.04-0.3)</td>
</tr>
<tr>
<td>PHQ≥10</td>
<td>136*</td>
<td>45</td>
<td>40</td>
<td>0</td>
<td>51</td>
<td>1(.90-1)</td>
<td>.56(.54-.66)</td>
<td>.53</td>
<td>1</td>
<td>2.3(1.8-2.9)</td>
<td>0 (np)</td>
</tr>
</tbody>
</table>

a: true positives; b: false positives; c: false negatives; d: true negatives; ppv, positive predictive value; npv, negative predictive value; LR, likelihood ratio.

*Missing (n=1) of which the sum score could not be calculated; np, not possible, the calculation cannot be performed because the values entered include one or more instances of zero.
Co-morbid anxiety or somatoform disorder

Of the 664 patients, 93 scored positive on the PHQ-9 anxiety and/or panic module. The adjusted prevalence of depressive disorder in this subgroup of 93 patients was 49.5%. The sensitivity and specificity figures for the diagnostic algorithm were 0.85 (95% CI: 0.71–0.93) and 0.83 (95% CI: 0.69–0.92), respectively, with a positive predictive value of 0.83, a negative predictive value of 0.85, a positive LR of 5.0 (95% CI: 2.6–9.5) and a negative LR of 0.2 (95% CI: 0.09–0.4). The sensitivity and specificity figures for the sum score of 10 and higher were 1 (95% CI: 0.90–1) and 0.45 (95% CI: 0.30–0.60), respectively, with a positive predictive value of 0.63, a negative predictive value of 1, a positive LR of 1.8 (95% CI: 1.4–2.3) and a negative LR of 0 (95% CI: not possible) (Table 2B).

Of the 664 patients, 137 scored positive on the PHQ-9 somatoform disorder module. The adjusted prevalence of depressive disorder in this subgroup of 137 patients was 33.6%. The sensitivity and specificity figures for the diagnostic algorithm were 0.91 (95% CI: 0.78–0.97) and 0.85 (95% CI: 0.75–0.91), respectively, with a positive predictive value of 0.75, a negative predictive value of 0.95, a positive LR of 5.9 (95% CI: 3.6–9.7) and a negative LR of 0.1 (95% CI: 0.04–0.3). The sensitivity and specificity figures for the sum score of 10 and higher were 1 (95% CI: 0.90–1) and 0.56 (95% CI: 0.45–0.66), respectively, with a positive predictive value of 0.53, a negative predictive value of 1, a positive LR of 2.3 (95% CI: 1.8–2.9) and a negative LR of 0 (95% CI: not possible) (Table 2B).

Severity measurement

We obtained the PHQ-9 and HDRS-17 scores of 49 patients with depressive disorder according to the SCID-I, with a maximum time interval of 14 days (mean, 6.7 days) between the PHQ-9 and the HDRS-17. We did not perform an HDRS-17 measurement of all 78 patients with depressive disorder according to the SCID because not all patients were willing to participate. Of these patients, 33 were female (67.3%). The mean age was 45.1 years. The scatter plot displays a diagram of PHQ-9 sum scores against HDRS-17 scores (Fig. 4). An analysis using Pearson correlation coefficient indicates a statistically significant linear relationship between the PHQ-9 sum score and the HDRS-17 score (r=.52, two-tailed Ps.01), which can be considered moderate. When calculating the coefficient of determination (r²=.27), we can conclude that the PHQ-9 sum score accounts for only 27% of the variability in HDRS-17 scores. Although the correlation was highly significant, it leaves 73% of the variability still to be accounted for by other variables.
Discussion

The sensitivity and specificity of the PHQ-9 in our study did not differ from previous accuracy studies of the PHQ-9 in primary care, although our study population was different. We can conclude that the widely used cutoff point of 10 indeed shows the highest sensitivity and specificity for screening purposes. The positive predictive value is 47%, while the chance of missing a depressive disorder with a negative score is 1% in our high-risk population. Thus, for screening purposes, the PHQ-9 proves to be a reliable instrument.

When applying the algorithmic cutoff point for diagnostic purposes in a high-risk population, the positive predictive value increases to 67% with a specificity of 95%, which is relatively high but insufficient for clinical purposes; 33% of all positive patients will be diagnosed erroneously as depressed. Therefore, we conclude that the PHQ-9 is not specific enough to be used as a diagnostic instrument in a population at high-risk. It can be used in general practice for case finding, but should always be followed by a formal diagnostic procedure. Positive and negative predictive values are both functions of the prevalence of depression and, as such, will change in samples with lower prevalence. Therefore, our conclusions are valid for similar high-risk populations;
however, in populations with a lower prevalence of depressive disorder, the positive predictive value of this instrument decreases.

As could be expected, the co-morbidity of depression and anxiety and/or somatization occurred frequently. A co-morbid anxiety or somatoform disorder increases sensitivity, but decreases the specificity of the PHQ-9 for depressive disorder. The positive predictive value increases not just due to the high prevalence of depressive disorder in this subgroup but also due to this overlapping symptomatology. For diagnostic purposes, the PHQ-9 performs better in subgroups with anxiety or somatoform symptoms, making it a useful instrument in clinical practice when dealing with patients with an anxiety or somatoform disorder.

The PHQ-9 performs modestly when used to measure the severity of a depressive disorder. The statistically significant agreement with our ‘reference standard’ (HDRS-17) was only weak and showed large inter individual variations. We have not analyzed the follow-up data of individual patients during treatment; therefore, we do not know whether the total score can be used as an individual follow-up measurement during treatment.

**Methodological strengths and limitations**

Our study was conducted in a large group of primary care patients using a standardized diagnostic instrument (SCID-I) and a severity instrument (HDRS-17) that had been validated extensively. The two-stage design proved efficient, without influencing statistical preciseness. Confidence intervals for specificity were rather small (as in previous studies), and we were able to include many positives and patients with scores around the cutoff values.³

Younger and male patients were less inclined to participate, probably due to less interest in this screening program for depression and lack of time because of a busier daily life. Patients belonging to the MHP group probably had already discussed their psychological problems with the GP during regular consultations and therefore rejected participation. The fact that the mean PHQ-9 scores and the distribution among high-risk groups of patients who participated and patients who gave informed consent but could not be reached by telephone did not differ endorses the fact that no relevant selection bias has occurred at this stage.

Our severity analysis showed that the PHQ-9 has only moderate agreement with the HDRS-17. The scatter plot shows a large variation around the regression line, which implies a wide confidence interval. This is an important limitation that has to be noted.
when using the PHQ-9 as a severity instrument. There are several explanations for this variation. First, the HDRS-17 method that is used to measure severity is quite specific and more detailed than the PHQ-9. The PHQ-9 only measures the severity of the nine depression criteria, and all nine criteria have the same weight. Compared to the PHQ-9, the HDRS-17 emphasizes sleeping problems and restlessness. Furthermore, anxiety symptoms, sexual problems and illness realization are part of the HDRS-17, not of the PHQ-9. We can conclude that the definition of severity according to the HDRS-17 is different from the definition of severity according to the PHQ-9. These differences within instruments could partly explain the variability within the scatter plot. Furthermore, the PHQ-9 is a self-fill-in questionnaire, while the HDRS-17 is an interview where the interviewer exercises judgment. This difference between the instruments could also partly explain the wideness of the confidence interval.

**Implications**

Our study is the first accuracy study of the PHQ-9 within specific high-risk groups for depressive disorder in general practice and the first study to account for psychiatric co-morbidity. The PHQ-9 proved to be appropriate for screening purposes, considering its high negative LR. Only one in every 100 patients with depressive disorder is not identified incorrectly. As a diagnostic instrument, the PHQ-9 has limitations. Although used in a high-risk group, the positive predictive value is too low for the instrument to be used in a clinical setting, as this will lead to many false positives. Epidemiological prevalence studies that use the PHQ-9 as only a diagnostic instrument should also take notice of these diagnostic limitations (false positives and false negatives) and correct their prevalence figures with the test characteristics of the PHQ-9. In a former systematic review, it has been shown that, in an unselected primary care population, the positive predictive value of the PHQ-9 is about 59%, while the negative predictive value is about 97%. To correct the prevalence figures, one should use the test characteristics of the PHQ that belong to that specific population. If, according to the PHQ-9, 200 out of 1000 patients have a positive score on the PHQ-9, then \((200 \times 0.59) + (800 \times 0.03) = 142\) patients are probably suffering from depressive disorder, resulting in a prevalence of \(142/1000 = 14.2\%\) instead of 20%.

Regarding the ability of the PHQ-9 to measure severity, we conclude that our study showed disappointing results. As our patient group was small, more research in larger populations is needed to assess inter-individual variation. Individual follow-up studies are necessary to explore the instrument’s intra-individual variation and sensitivity to change. Until then, the PHQ-9 does not seem appropriate as a severity index.
Funding
Funding for this study was provided by a grant from the Netherlands Organization for Health Research and Development (ZonMw), program Mental Health (# 100.003.005 and # 100.002.021).

Acknowledgements
We thank the participants, Judith Kosman and the participating GPs for their cooperation.
REFERENCES


