Which instruments to support diagnosis of depression have sufficient accuracy?
A systematic review

AGNETA PETTERSSON, KRISTINA BENGTSSON BOSTRÖM, PETTER GUSTAVSSON, LISA EKSELIUS


Background: Instruments are frequently used in case finding, diagnosis and severity grading of major depression, but the evidence supporting their utility is weak. Aim: To systematically review the specificity and sensitivity of instruments used to diagnose and grade the severity of depression. Methods: MEDLINE, PsychInfo, Embase and the Cochrane Library databases were searched until April 2014. Fifty studies fulfilled the inclusion criteria. Risk of bias was assessed with QUADAS. The average sensitivity and specificity of each instrument was estimated with hierarchical summary receiver operating characteristics analyses and the confidence in the estimates was evaluated using GRADE. Minimum acceptable sensitivity/specificity, with structured interview as the reference, was 80%/80% for structured interviews and 80%/70% for case-finding instruments. The minimum acceptable standard for severity measures was a correlation of 0.7 with DSM-IV classification. Results: Twenty instruments were investigated. The average sensitivity/specificity was 85%/92% for the Structured Clinical Interview for DSM-IV-Axis I Disorders (SCID-I), 95%/84% for the Mini International Neuropsychiatric Interview (MINI), <70%/85% for the Primary Care Evaluation of Mental Disorders (PRIME-MD), 88%/78% for the Patient Health Questionnaire-9 (PHQ-9) with a cut-off score of 10, 69%/95% for PHQ-9 as a diagnostic algorithm and 70%/83% for the Hospital Anxiety and Depression Scale (HADS) with a cut-off score of 7. The confidence in the estimates for the other instruments was very low. Conclusions: Only the SCID-I, MINI and PHQ-9 with a cut-off score of 10 fulfilled the minimum criteria for sensitivity and specificity. The use of the PRIME-MD and HADS is not supported by current evidence.

• Depressive disorder, Diagnosis, Evidence, Interview, Meta-analysis, Questionnaires, Psychiatric status rating scales, Psychological standards, Standards.

Agneta Pettersson, SBU, PO 3657, 103 59 Stockholm, Sweden, E-mail: pettersson@sbu.se; Accepted 13 January 2015.

Case finding and diagnosis are critical in the management of depression. Patients who do not have a diagnosis in the medical record system have a smaller chance of being managed according to guidelines (1). A structured diagnostic procedure would ideally be based on a longitudinal assessment of all available data, such as the LEAD (Longitudinal Expert Assessment All Data) procedure or Best Estimate procedure (2, 3). However, in routine practice, diagnosis most often relies on information gathered at the consultation. A systematic review on the diagnostic accuracy of unassisted diagnoses of depression highlighted the weaknesses of this approach (4). General practitioners correctly ruled out depression in most patients but correctly identified depression in less than half of the patients (4). Similar results have been obtained for psychiatric outpatient clinics (5), but correct identification of depression is higher in hospitalized patients, at over 80% (6, 7).

Many instruments have been developed to support a structured collection of information for case finding, diagnosis and severity grading of depression. However, the patient benefits of these instruments have not been ascertained and the literature shows conflicting results (8–11). While awaiting a consensus on patient benefits, instruments that have sufficiently high diagnostic accuracy to be of value would be preferred. There is no agreement in the literature on the minimal requirements (“benchmark”) for diagnostic accuracy of instruments for depression, although an earlier systematic review suggested a sensitivity of at least 85% and a specificity of at least 75% for case-finding instruments (12).
There are recent systematic reviews of the diagnostic accuracy instruments for case finding (13–15). Meader et al. (14) evaluated 13 case-finding instruments for depression in patients with chronic illnesses, Manea et al. (15) assessed the sensitivity and specificity of the Hospital Anxiety and Depression Scale (HADS). The sensitivity and specificity of PHQ-9 at the established cut-off score of 10 were 82–84% and 88–89%, respectively, and the sensitivity and specificity of HADS at a cut-off score of 7–8 were 75–82% and 74–81%, respectively. However, these systematic reviews did not account for study bias due to methodological weaknesses. Bias due to deficiencies in blinding, verification and sampling method have been shown to give inflated results (16). Furthermore, the perception of depression may vary depending on cultural factors, which may affect responses to instruments based on the Diagnostic and Statistical Manual (DSM) system (17–19). The aims of this systematic review were to assess the accuracy of case-finding instruments, structured interviews and severity measures for major depression in adults in clinical populations and to determine which instruments met a benchmark level of diagnostic accuracy.

**Materials and methods**

This systematic review is an update of a report that was commissioned by the Swedish Ministry of Health and Social Affairs to the Swedish Council on Health Technology Assessment and published in Swedish in 2012 (20).

**Inclusion criteria**

Studies of patients with symptoms of depression according to DSM-III, DSM-III-R, DSM-IV or the International Classification of Diseases (ICD-9, ICD-10) or a high risk of depression were considered for inclusion in this systematic review. Settings were confined to Europe, North America, Australia and New Zealand to minimize the effect of cultural differences on outcomes. The index tests were 20 instruments that, according to an open-ended questionnaire addressed to the Swedish Psychiatric Association and the Swedish College of General Practice, were used routinely. To ensure that the results would be applicable to routine practice, the performance of the instruments was evaluated using the established cut-off value.

The LEAD procedure was considered the gold standard for the diagnosis of depression. In the lack of an exact standardized measure against which to determine the validity of psychiatric diagnoses, LEAD, “the longitudinal, expert, all data procedure”, has been suggested to represent the most appropriate strategy for defining a gold standard (2). In a LEAD procedure, expert clinicians utilize all the available data over time, including information from family members, hospital records, psychological evaluation, and laboratory results as a basis for a diagnosis. As the number of studies using LEAD as the gold standard was limited we also accepted studies using any structured interview based on DSM or ICD classifications as a reference standard for structured interviews or case-finding instruments. For severity measures, the DSM-IV classification of severity (21) was used, operationalized by the Structured Clinical Interview for DSM-IV-Axis I (SCID-I) and the Clinical Global Impression of Severity (CGI-S), and these instruments were chosen as reference standards. Blinding was required for severity studies, i.e. the same rater should not perform both the index test and SCID-I (or CGI-S). For case finding and diagnosis of depression, the time between the index and reference tests should be one week or less, and for severity measurement 24 h or less. The outcome of studies of structured interviews and case-finding instruments was the sensitivity and specificity of the index test for detecting major depression. The studies should report sufficient information to allow us to construct 2 × 2 tables.

The performance of an instrument was judged against a minimum acceptable sensitivity and specificity or correlation coefficient, respectively. We used 80% for sensitivity and 70% for specificity for case-finding instruments, 80% for both sensitivity and specificity for structured interviews, and a correlation coefficient of 0.70 for instruments to assess depression severity.

**Literature search**

Literature searches were conducted in PubMed, PsycInfo, Embase and the Cochrane Library databases up to May 2011. The search in the PubMed database was updated, covering the time from May 2011 to April 29, 2014 using the same terms as in the original search. The search strategies are available at www.sbu.se/affective_disorders. Full-text articles published in peer-reviewed journals in the English, Scandinavian, German and French languages were considered for inclusion. Studies cited in the reference lists of included studies were also searched. The systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (22). Researchers worked in pairs and independently selected studies from lists of abstracts, assessed the relevance and risk of bias of each study, and extracted data from each study. In the case of a disagreement, the study was processed by the entire research group. The risk of bias was assessed using the QUADAS checklist (23) with an extra item on adequate education of the raters regarding use of the reference standard. The researchers were trained using a crib sheet until the agreement reached an acceptable level (Fleiss’ kappa > 0.7) (24). The risk of bias per
item in the QUADAS checklist and per study was entered in a risk-of-bias table (25). Only studies with a low or moderate risk of bias were included in the main analyses. Studies with a high risk of bias were used in subgroup analyses.

Statistics
The average sensitivity and specificity of each instrument were calculated using hierarchical summary receiver operating characteristics (HSROC) analyses, as outlined in the Cochrane Handbook (26). Values for true and false positives and negatives for individual studies were entered in a macro (METADAS) based on the Rutter-Gatsonis HSROC model (27) in SAS software Version 9.3. Data from METADAS were exported to RevMan 5.2, where the HSROC plots that show sensitivity and specificity for individual studies, as well as the average and 95% confidence region of sensitivity and specificity were visualized. For instruments that were assessed in only two studies, data were entered in a fixed effects model (26) in MetaDisc software (28).

Assessing confidence in the estimate
The confidence in the average sensitivity and specificity of each instrument was evaluated using the GRADE methodology (29), which classifies the confidence as high, moderate, low or very low (Table 1). The preliminary confidence was set as high (⊕⊕⊕⊕). Thereafter, the confidence was judged with respect to five domains: overall risk of bias, heterogeneity between studies (inconsistency), the width of the CI for the estimate (imprecision), problems with applicability (indirectness) and the risk of publication bias. The confidence was lowered if there were deficiencies in one or several of these domains. A priori, it was decided that imprecision should be related to the minimum acceptable benchmark criteria. If the 95% CI of the average crossed the benchmark, the confidence in the precision was reduced by one level.

<table>
<thead>
<tr>
<th>Confidence in estimate</th>
<th>Characteristics</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>High (⊕⊕⊕⊕)</td>
<td>Only minor problems across the five domains (risk of bias, inconsistency, imprecision, indirectness and publication bias)</td>
<td>High confidence in the estimate, which with high probability will not be changed by new research</td>
</tr>
<tr>
<td>Moderate (⊕⊕⊕)</td>
<td>Severe problem in one domain</td>
<td>The estimate is likely to be true</td>
</tr>
<tr>
<td>Low (⊕⊕)</td>
<td>Very severe problems in one domain or severe problems in two domains or minor problems in several domains</td>
<td>The estimate is probably true but may be changed by new research</td>
</tr>
<tr>
<td>Very low (⊕)</td>
<td>Severe or very severe problems in several domains</td>
<td>Little confidence in the estimate. High probability that the estimate will be changed by the next published study</td>
</tr>
</tbody>
</table>

Results
The literature search in the original report generated 33,224 abstracts across patients of all ages and diagnoses. Restricting the search to “adults” and “depression” resulted in 483 articles that were read in full. Forty-three of these fulfilled the inclusion criteria (Fig. 1). The updated search generated 986 new abstracts, whereof 18 articles were read in full text and seven were included. Excluded studies are listed at www.sbu.se/affective_disorders. Fifteen of the studies had a high risk of bias and were excluded from the meta-analyses. The results described below are from the 35 studies with a low or moderate risk of bias.

Structured interviews for diagnosis of depression
Eight structured interviews were investigated (Table 2). The SCID-I, the Mini International Neuropsychiatric Interview (MINI) (30) and the Primary Care Evaluation of Mental Disorders (PRIME-MD) (31) were assessed in two studies each, all with a moderate risk of bias. The estimates of sensitivity and specificity from the fixed model meta-analyses and the confidence in the estimates are shown in Table 3. SCID-I was compared with LEAD in both studies and had a sensitivity of 86%. The lower boundary for the 95% CI was 73%; therefore, the confidence was lowered one level due to imprecision. The MINI had a sensitivity of 95% and a specificity of 84% and the confidence in the average was high. The sensitivity of the PRIME-MD differed widely between the two studies (30% and 68%); therefore, the values were not averaged to a point estimate. However, we concluded that average sensitivity was below 70%.

The Composite International Diagnostic Interview (CIDI) (32) and the Diagnostic Interview Schedule (33) were each evaluated in one study with a moderate risk of bias (Table 2), and the Schedule for Affective Disorders and Schizophrenia (SADS) (34) was evaluated in one study with a high risk of bias. The quality of the evidence for these three structured interviews was very
low (Table 3). No studies on the Schedules for Clinical Assessment in Neuropsychiatry (35) or the Structured Psychiatric Interview for General Practice (36) fulfilled the inclusion criteria.

We concluded that only SCID-I and MINI fulfilled the benchmark levels.

Case-finding instruments
Six case-finding instruments were investigated (Table 4). No studies on the Calgary Depression Scale for Schizophrenia (CDSS) (37) or the Major Depression Inventory (38) fulfilled the inclusion criteria. The Beck Depression Inventory (BDI-II) (39) at a cut-off score of 14, the Center for Epidemiologic Studies of Depression Scale (CES-D) (39, 40) at a cut-off score of 16, the depression subscale of the HADS (41) at a cut-off score of 7 and the PHQ-9 (42) at a cut-off score of 10 were evaluated with HSROC analysis. The PHQ-9 is also used as a diagnostic algorithm instead of a cut-off score, and this algorithm-based PHQ-9 was also assessed. The estimates of sensitivity and specificity and the confidence in the estimates are presented in Table 5.

The BDI-II was evaluated in four studies that were included in the meta-analysis. They included a total of 824 patients from chronic pain centres (43), clinics for advanced cancer (44), rehabilitation after acute cardiac events (45) and primary care settings (46). The sensitivity of the BDI-II was consistent across studies, between 88% and 100%, but the specificity varied from 50% to 84%. The average sensitivity was 92% and the average specificity was 72%. However, the confidence in these estimates was low as data could not be fitted to the model and the average was an approximation. A subgroup analysis to which six studies (47–52) with a high

low (Table 3). No studies on the Schedules for Clinical Assessment in Neuropsychiatry (35) or the Structured Psychiatric Interview for General Practice (36) fulfilled the inclusion criteria.

We concluded that only SCID-I and MINI fulfilled the benchmark levels.

Case-finding instruments
Six case-finding instruments were investigated (Table 4). No studies on the Calgary Depression Scale for Schizophrenia (CDSS) (37) or the Major Depression Inventory (38) fulfilled the inclusion criteria. The Beck Depression Inventory (BDI-II) (39) at a cut-off score of 14, the Center for Epidemiologic Studies of Depression Scale (CES-D) (39, 40) at a cut-off score of 16, the depression subscale of the HADS (41) at a cut-off score of 7 and the PHQ-9 (42) at a cut-off score of 10 were evaluated with HSROC analysis. The PHQ-9 is also used as a diagnostic algorithm instead of a cut-off score, and this algorithm-based PHQ-9 was also assessed. The estimates of sensitivity and specificity and the confidence in the estimates are presented in Table 5.

The BDI-II was evaluated in four studies that were included in the meta-analysis. They included a total of 824 patients from chronic pain centres (43), clinics for advanced cancer (44), rehabilitation after acute cardiac events (45) and primary care settings (46). The sensitivity of the BDI-II was consistent across studies, between 88% and 100%, but the specificity varied from 50% to 84%. The average sensitivity was 92% and the average specificity was 72%. However, the confidence in these estimates was low as data could not be fitted to the model and the average was an approximation. A subgroup analysis to which six studies (47–52) with a high

Table 2. Sensitivity and specificity of structured interviews for diagnosing major depression.

<table>
<thead>
<tr>
<th>Study</th>
<th>Patients</th>
<th>Index test</th>
<th>Reference test</th>
<th>Sensitivity and specificity</th>
<th>Risk of bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miller, 2001 (6)</td>
<td>n = 75 (psychiatry, inpatients)</td>
<td>SCID-I</td>
<td>LEAD</td>
<td>Sensitivity: 92%; specificity: 98%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Ramirez-Basco, 2001 (5)</td>
<td>n = 210 (psychiatry, outpatients)</td>
<td>SCID-I</td>
<td>LEAD</td>
<td>Sensitivity: 84%; specificity: 91%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Lecrubier, 1997 (92)</td>
<td>n = 350 (psychiatry, outpatients)</td>
<td>MINI</td>
<td>CIDI</td>
<td>Sensitivity: 94%; specificity: 79%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Sheehan, 1997 (93)</td>
<td>n = 320 (psychiatry, outpatients)</td>
<td>MINI</td>
<td>SCID-P</td>
<td>Sensitivity: 96%; specificity: 88%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Leopold, 1998 (94)</td>
<td>n = 122 (oncology)</td>
<td>PRIME-MD</td>
<td>SCID-I</td>
<td>Sensitivity: 30%; specificity: 93%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Loerch, 2000 (95)</td>
<td>n = 924 (PCCs and psychiatry, outpatients)</td>
<td>PRIME-MD</td>
<td>M-CIDI</td>
<td>Sensitivity: 68%; specificity: 84%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Booth, 1998 (96)</td>
<td>n = 54 (hospitalized for somatic reasons)</td>
<td>CIDI</td>
<td>LEAD</td>
<td>Sensitivity: 67%; specificity: 84%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Hasin, 1987 (97)</td>
<td>n = 120 (alcohol rehabilitation clinic)</td>
<td>DIS</td>
<td>SADS</td>
<td>Sensitivity: 25%; specificity: 90%</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

PCC, primary care centres.
risk of bias were added supported the finding that the BDI-II has a high sensitivity and a specificity that varies between studies.

The CES-D was evaluated in two studies with a moderate risk of bias. One study was performed in patients at a rehabilitation clinic (53) and one assessed mothers of disabled children (54). In the fixed effects model, the pooled sensitivity was 95%. Specificity was 33% in one study and 73% in the other study and was not pooled. The confidence in the estimates was very low (Table 5).

The HADS was evaluated in 10 studies with a low or moderate risk of bias (Table 4). The total number of patients was 4928. Four studies were performed in outpatient clinics at hospitals (55–58), one was performed in an emergency ward (59), three in a primary care setting (60–62) and two involved patients at follow-up aftercare at specialist clinics (63, 64). The average sensitivity was 70% and the average specificity was 83% (Fig. 2). As with the BDI-II, the average point was an approximation. Analysis showed that the two studies on patients with heart disease (62, 64) did not fit into the model. After these studies were removed, the sensitivity was 75% and the specificity was 81%. However, as there was no clear reason for excluding these studies, they were reinserted. The quality of evidence for the estimates was moderate (Table 5).

The PHQ-9 based on a diagnostic algorithm was investigated in 11 studies with a low or moderate risk of bias from a broad range of settings in Europe, the USA and Australia (Table 4). They included a total of 9990 patients and seven of them used SCID-I as reference standard. Five studies were performed in a primary care setting (42, 61, 62, 65, 66), four in somatic outpatient clinics (64, 67–69) and two in somatic inpatient wards (70, 71). The HSROC analysis revealed an average sensitivity of 69% and an average specificity of 95%. The size of the confidence region lowered the confidence for sensitivity to moderate, but did not affect the high confidence for specificity (Table 5).

PHQ-9 with a cut-off score of 10 was evaluated in 10 studies with a low or moderate risk of bias that included a total of 9517 patients and used a variety of reference standards (Table 4). Three studies were performed in a primary care setting (61, 65, 72), six in an outpatient clinic (58, 62, 64, 67, 73, 74) and one in a trauma unit for spinal cord injuries (75). The HSROC analysis revealed an average sensitivity of 88% and an average specificity of 78% with a large confidence region (Fig. 3). The confidence in the evidence was moderate for both sensitivity and specificity (Table 5).

We concluded that only PHQ-9 at a cut-off of 10 fulfilled benchmark levels.

### Instruments for severity measurement

Seven instruments, CDSS, Hamilton Depression Rating Scale-17 items (HDRS-17) (76), Inventory of Depressive Symptomatology (IDS) (77), Montgomery Asberg Depression Rating Scale (MADRS) (78), PHQ-9, BDI-II and the Zung Self-Rated Depression Scale (79) were investigated. Of 105 studies read in full, the vast majority compared two depression scales against another and those were excluded. Five studies on HDRS-17 were included but only one had an acceptable risk of bias. This study used a revised version of the HDRS-17 (80) to grade the severity of depression in 47 elderly inpatients with dementia and 98 relatives (81). The agreement between the HDRS-17 score and the CGI-S score was evaluated using Spearman rank correlation and was 0.85. The confidence in this estimate was very low (Table 6).
Table 4. Sensitivity and specificity of case-finding instruments for major depression with structured interviews as reference standard.

<table>
<thead>
<tr>
<th>Study</th>
<th>Patients</th>
<th>Index test</th>
<th>Reference test</th>
<th>Sensitivity and specificity</th>
<th>Risk of bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buvencius, 2012 (45)</td>
<td>n = 522 (cardiac rehabilitation clinic)</td>
<td>BDI-II</td>
<td>MINI</td>
<td>Sensitivity: 89%; specificity: 74%</td>
<td>Low</td>
</tr>
<tr>
<td>Dutton, 2004 (46)</td>
<td>n = 223 (PCC waiting room)</td>
<td>BDI-II</td>
<td>PRIME-MD</td>
<td>Sensitivity: 88%; specificity: 84%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Poole, 2009 (43)</td>
<td>n = 36 (pain specialist centre)</td>
<td>BDI-II</td>
<td>SCID-I</td>
<td>Sensitivity: 100%; specificity: 50%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Warmenhoven, 2011 (44)</td>
<td>n = 61 (oncology clinic)</td>
<td>BDI-II</td>
<td>PRIME-MD</td>
<td>Sensitivity: 90%; specificity: 64%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Hall, 1999 (57)</td>
<td>n = 269 (breast cancer outpatients)</td>
<td>HADS PSE</td>
<td></td>
<td>Sensitivity: 33%; specificity: 93%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Haddad, 2013 (62)</td>
<td>n = 730 (cardiac patients, PCCs)</td>
<td>HADS CIS-R</td>
<td></td>
<td>Sensitivity: 53%; specificity: 91%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Lowe, 2004 (61)</td>
<td>n = 2050 (PCCs and hospital outpatients)</td>
<td>HADS SCID-I</td>
<td></td>
<td>Sensitivity: 88%; specificity: 69%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Orive, 2010 (58)</td>
<td>n = 167 (hospital waiting rooms)</td>
<td>HADS PRIME-MD</td>
<td></td>
<td>Sensitivity: 86%; specificity: 75%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Silverstone, 1994 (59)</td>
<td>n = 189 (7 days after acute hospitalization)</td>
<td>HADS SCAN</td>
<td></td>
<td>Sensitivity: 100%; specificity: 73%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Stafford, 2007 (64)</td>
<td>n = 528 (cardiac outpatients)</td>
<td>HADS MINI</td>
<td></td>
<td>Sensitivity: 46%; specificity: 92%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Sultan, 2010 (99)</td>
<td>n = 370 (diabetes clinic)</td>
<td>HADS MINI</td>
<td></td>
<td>Sensitivity: 53%; specificity: 86%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Terlui, 2009 (60)</td>
<td>n = 370 (on sick leave, PCCs)</td>
<td>HADS CIDI</td>
<td></td>
<td>Sensitivity: 93%; specificity: 39%</td>
<td>Low</td>
</tr>
<tr>
<td>Zeger, 2004 (63)</td>
<td>n = 98 (tinnitus outpatients)</td>
<td>HADS SCID-I</td>
<td></td>
<td>Sensitivity: 80%; specificity: 94%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Bombardier, 2012 (75)</td>
<td>n = 142 (spinal cord injuries)</td>
<td>PHQ-9</td>
<td>SCID-I</td>
<td>Cut-off score of 10: Sensitivity: 100%; specificity: 80%</td>
<td>Low</td>
</tr>
<tr>
<td>Cassin, 2013 (73)</td>
<td>n = 244 (bariatric surgery candidates)</td>
<td>PHQ-9</td>
<td>MINI</td>
<td>Cut-off score of 10: Sensitivity: 75%; specificity: 63%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Cassin, 2013 (73)</td>
<td>n = 275 (bariatric surgery candidates)</td>
<td>PHQ-9</td>
<td>MINI</td>
<td>Cut-off score of 10: Sensitivity: 80%; specificity: 46%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Diez-Quevedo (71)</td>
<td>n = 1003 (medicinal and surgical inpatients)</td>
<td>PHQ-9</td>
<td>SCID-I</td>
<td>Algorithm: Sensitivity: 84%; specificity: 92%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Elderon, 2011 (74)</td>
<td>n = 1024 (cardiac outpatients)</td>
<td>PHQ-9</td>
<td>C-DIS</td>
<td>Cut-off score of 10: Sensitivity: 54%; specificity: 90%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Haddad, 2013 (62)</td>
<td>n = 730 (cardiac outpatients, PCCs)</td>
<td>PHQ-9</td>
<td>CIS-R</td>
<td>Algorithm: Sensitivity: 59%; specificity: 95%; Cut-off score of 10: Sensitivity: 84%; specificity: 90%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Henkel, 2004 (66)</td>
<td>n = 470 (PCCs)</td>
<td>PHQ-9</td>
<td>CIDI</td>
<td>Algorithm: Sensitivity: 78%; specificity: 85%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Lowe, 2004 (61)</td>
<td>n = 2050 (outpatients medical hospital and PCCs)</td>
<td>PHQ-9</td>
<td>SCID-I</td>
<td>Algorithm: Sensitivity: 83%; specificity: 90%; Cut-off score of 11: Sensitivity: 90%; specificity: 77%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Navines, 2012 (69)</td>
<td>n = 500 (HCV outpatients)</td>
<td>PHQ-9</td>
<td>SCID-I</td>
<td>Algorithm: Sensitivity: 72%; specificity: 99%</td>
<td>Low</td>
</tr>
<tr>
<td>Orive, 2010 (58)</td>
<td>n = 53 (hospital waiting rooms)</td>
<td>PHQ-9</td>
<td>PRIME-MD</td>
<td>Cut-off score of 10: Sensitivity: 68%; specificity: 89%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Persoons, 2003 (68)</td>
<td>n = 97 (otolaryngology outpatients)</td>
<td>PHQ-9</td>
<td>MINI</td>
<td>Algorithm: Sensitivity: 68.8%; specificity: 94.4%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Picardi, 2005 (70)</td>
<td>n = 141 (dermatology, inpatients)</td>
<td>PHQ-9</td>
<td>SCID-I</td>
<td>Algorithm: Sensitivity: 55%; specificity: 91%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Spitzer, 1999 (42), Kroenke, 2001 (72)</td>
<td>n = 3000 (PCCs)</td>
<td>PHQ-9</td>
<td>SCID-I</td>
<td>Algorithm: Sensitivity: 73%; specificity: 98%; Cut-off score of 10: Sensitivity: 88%; specificity: 88%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Stafford, 2007 (64)</td>
<td>n = 528 (cardiac outpatients)</td>
<td>PHQ-9</td>
<td>MINI</td>
<td>Algorithm: Sensitivity: 34%; specificity: 97%; Cut-off score of 10: Sensitivity: 54%; specificity: 91%</td>
<td>Moderate</td>
</tr>
<tr>
<td>Thekkumpurath, 2011 (67)</td>
<td>n = 782 (regional oncology clinics)</td>
<td>PHQ-9</td>
<td>SCID-I</td>
<td>Algorithm: Sensitivity: 56%; specificity: 96%; Cut-off score of 10: Sensitivity: 73%; specificity: 88%</td>
<td>Moderate</td>
</tr>
</tbody>
</table>
INSTRUMENTS TO SUPPORT DIAGNOSIS OF DEPRESSION

No studies with an acceptable risk of bias were retrieved for the other instruments.

Conclusions
Only three out of 20 assessed instruments fulfilled our benchmark criteria: two structured interviews, SCID-I and MINI, and one case-finding instrument, PHQ-9 with a cut-off score of 10. The structured interview PRIME-MD and the case-finding instruments HADS with a cut-off score of 7 and PHQ-9 as a diagnostic algorithm had sensitivities that were too low to be useful in clinical practice. BDI-II with a cut-off score of 14 had adequate sensitivity but low specificity. No severity measures were supported by evidence despite a large number of studies.

Although standardized interviews are the gold standard for the diagnosis of depression in clinical research (82), remarkably few studies have scrutinized their sensitivity and specificity. Their accuracy seems to have been taken for granted. Only three structured interviews, SCID-I, SADS and CIDI, have been validated against the LEAD procedure, which can be considered as the best available reference standard. Furthermore, accepting any structured interview as the reference standard, no structured interview was supported by more than two studies with a low or moderate risk of bias. MINI has been evaluated in two other studies with a high risk of bias (83, 84). Both showed a sensitivity above 80%, ranging from 85% to 100%, which adds further support to our finding. For PRIME-MD, one study by Spitzer et al. was excluded due to the time interval between tests (85). However, that study found a sensitivity of 57%, which is in line with our assessment.

Table 5. Summary of findings for diagnostic accuracy of case-finding instruments for major depression with structured interviews as reference standard.

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Outcome</th>
<th>Studies (n); patients (n)</th>
<th>Confidence in estimate</th>
<th>Reasons for downgrading confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDI-II, cut-off score of 14</td>
<td>Sensitivity</td>
<td>4; 824</td>
<td>92% (83–97%)</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>4; 824</td>
<td>72% (58–82%)</td>
<td>Very low</td>
</tr>
<tr>
<td>HADS, cut-off score of 7</td>
<td>Sensitivity</td>
<td>10; 4928</td>
<td>70% (55–82%)</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>10; 4928</td>
<td>83% (73–90%)</td>
<td>Low</td>
</tr>
<tr>
<td>PHQ-9, algorithm-based</td>
<td>Sensitivity</td>
<td>11; 9990</td>
<td>69% (60–76%)</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>11; 9990</td>
<td>95% (92–97%)</td>
<td>High</td>
</tr>
<tr>
<td>PHQ-9, cut-off score of 10</td>
<td>Sensitivity</td>
<td>10; 9517</td>
<td>88% (77–94%)</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>10; 9517</td>
<td>78% (65–88%)</td>
<td>Moderate</td>
</tr>
<tr>
<td>CES-D, cut-off score of 16</td>
<td>Sensitivity</td>
<td>2; 459</td>
<td>95% (83–99%)</td>
<td>Very low</td>
</tr>
<tr>
<td></td>
<td>Specificity</td>
<td>2; 459</td>
<td>33–73%</td>
<td>Very low</td>
</tr>
</tbody>
</table>

*Deficiencies in meta-analysis.
†Heterogeneous studies.
‡Lower 95% CI below benchmark.
§Six out of 10 studies outside 95% confidence region.
||Blinding unclear.
¶Narrow patient spectrum.
Our estimates for PHQ-9 with a cut-off score of 10 are in line with the systematic reviews by Meader et al. (14) and Manea et al. (15), although we found somewhat higher sensitivity and lower specificity than they did. It has to be stressed that PHQ-9 only is supported with evidence as a screening (i.e. case finding) instrument. There is no evidence for PHQ-9 as a severity measure and thus as an aid to monitor treatment effects.

Our results for HADS must be discussed. First, the sensitivity in the present review is lower than reported by Meader et al. (14) and by Brennan et al. (13). The difference may be an artefact, as the three systematic reviews used different inclusion and quality criteria. Second, some studies did not fit into the Rutter–Gatsonis model that we used in our meta-analysis so the average sensitivity and specificity were only approximations. This is in line with current discussions questioning the transferability and the factor structure of HADS. In a recent study, Maters et al. (86) argued that discrepancies between studies may depend on difficulties in translating HADS from the original, everyday British-English language to other languages without loss of the intended meaning. According to Maters et al., the effect of inexact translations could be a change in the optimal cut-off level. However, with only 10 studies included in our analysis, no conclusions could be drawn regarding bias due to translation issues. Several authors using Item Response Theory to study differential item functioning

---

**Table 6.** Summary of findings for severity measures for major depression with Clinical Global Impression-Severity (CGI-S) as reference standard.

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Outcome</th>
<th>Studies (n); patients (n)</th>
<th>Summary estimate</th>
<th>Confidence in estimate</th>
<th>Reasons for grading down confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDRS-17</td>
<td>Correlation with CGI-S</td>
<td>1 (98)</td>
<td>Spearman rank correlation $r = 0.85$</td>
<td>Very low</td>
<td>Indirectness: $-2^*$; Imprecision: $-1^+$</td>
</tr>
</tbody>
</table>

*Few patients, narrow patient spectrum.
†One study.
related to demographics (87–89) have shown that gender, age and disease impact the sensitivity and specificity of HADS. This may be the reason that the two single studies on populations with coronary diseases did not fit into our model and that the sensitivity ranged from 33% to 100% and specificity from 39% to 94%.

A systematic review by Wang & Gorenstein (90) identified five studies that compared the BDI-II with a structured interview at a cut-off of 14. In this review, no meta-analysis was performed and the risk of bias was not considered, but the sensitivity ranged from 88% to 94% and the specificity ranged from 74% to 84%, which is higher than what we found.

A striking finding was the absence of evidence supporting instruments designed to measure depression severity. Although there was a large body of studies that investigated these instruments, the vast majority compared the change in the scores of two or more instruments during treatment for depression. Although these studies showed that the instruments follow each other in the same direction, they did not anchor the severity scores to the DSM classification mild, moderate and severe.

The results of this systematic review are influenced by the strengths and limitations of our methodology. We consider our choice to follow the PRISMA guidelines (22), to include only studies with low or moderate risk of bias in the analysis, and to examine the confidence in the estimates with GRADE as strengths of the review.

A risk of bias may be introduced if raters using structured or semi-structured interviews are not adequately trained. This risk was assessed by an extra item in QUADAS as described in the Methods section. Most studies specified the education and training of the raters. Raters were commonly trained psychologists or psychiatrists and we perceive that the risk of bias due to insufficient training is small.

In retrospect, our definition of reference standard was a weakness in this review. LEAD and SCID-I were seldom used in studies. We therefore chose to accept a minor degree of uncertainty by allowing diagnosis by any structured interview based on DSM or ICD classification as a reference. This may have introduced bias into our results. When summarizing our results we for example realized that the PRIME-MD was the reference standard in three included studies (44, 46, 58). Two of them (44, 46) evaluated the BDI-II and gave results similar to those of the other two included studies. The third (58) investigated HADS and PHQ-9 and contributed less than 5% of the total sample size to the meta-analyses for these instruments. The studies thus had little impact on the results.

A limitation is that the setting was confined to a number of developed countries and studies from other parts of the world were excluded. This was done to minimize the risk that obtained results would be flawed by cultural factors, as shown for example in (91). However, this may limit the generalizability of our results.

A potential limitation in systematic reviews is that results are skewed due to conflicts of interest and publication bias. When an instrument is connected with a licence fee, certifications for use or copyright there is a risk that unfavourable studies are not published. When a stakeholder has copyrights on an instrument, the unlimited preparedness to allow for its use by independent researchers is necessary to avoid such an effect. This latter situation is applicable for the PHQ-9 instrument, which is copyrighted by a pharmaceutical company. It has, however, been evaluated in a large number of studies without financial support from the company in question, why there is no reason to believe that our estimates are affected by selective reporting.

This systematic review offers practical advice to clinicians who want to improve their diagnostic accuracy of major depression with the use of evidence-based instruments. A high level of diagnostic accuracy is crucial in clinical practice, and without it adequate treatment cannot be given. It also constitutes the basis for both treatment studies and studies on the aetiology, epidemiology and pathophysiology of disease. The use of instruments with unsatisfactory diagnostic accuracy casts doubt on the results of such studies.

Acknowledgments—We thank Cecilia Svanborg, M.D., Ph.D. and Professor Maria Tillfors, who assessed the structured interviews and BDI-II, Professor Måns Rosén, who gave valuable comments throughout the manuscript process, Derya Akcan and Agneta Bruland, who performed the literature search, and Lars Berglund at Uppsala Clinical Research Center, who performed the HSROC calculations.

Disclosure of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References


34. Endicott J, Spitzer RL. A diagnostic interview. The schedule for affective disorders and schizophrenia. Arch General Psychiatry 1978;35:837–44.

35. Wing JK, Babor T, Brugha T, Burke J, Cooper JE, Giel R, et al. SCAN. Schedules for Clinical Assessment in Neuropsychiatry. Arch Gen Psychiatry 1990;47:589–93.


56. Whelan-Goodinson R, Ponsford J, Schonberger M. Validity of the Hospital Anxiety and Depression Scale to assess depression and anxiety following traumatic brain injury as compared with the Structured Clinical Interview for DSM-IV. J Affect Disord 2009.


60. Terluin B, Brouwers EP, van Marwijk HW, Verhaak PF, van der Horst HE. Detecting depressive and anxiety disorders in distressed patients in primary care; comparative diagnostic accuracy of the Four-Dimensional Symptom Questionnaire (4DSQ) and the Hospital Anxiety and Depression Scale (HADS). BMC Fam Pract 2009;10:58.


Kristina Bengtsson Boström, M.D., Ph.D., Research & Development Centre Skaraborg Primary Care, Skövde, and Department of Clinical Sciences/Endocrinology, Lund University, Malmö, Sweden.

Petter Gustavsson, Ph.D., Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden.

Ekstius Lisa, M.D., Ph.D., Department of Neuroscience, Uppsala University, Uppsala, Sweden.