informa
healthcare

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

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

Pettersson A, Bengtsson Boström K, Gustavsson P, Ekselius L. Which instruments to support diagnosis of depression have sufficient accuracy? A systematic review. Nord J Psychiatry 2015;Early Online:1–12.

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, PsycInfo, 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).

DOI: 10.3109/08039488.2015.1008568

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 Patient Health Questionnaire-9 (PHQ-9) and Brennan et al. (13) the properties 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-IIIR, 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.

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

Table 1. The GRADE classification (29) and our interpretation of the grading.

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

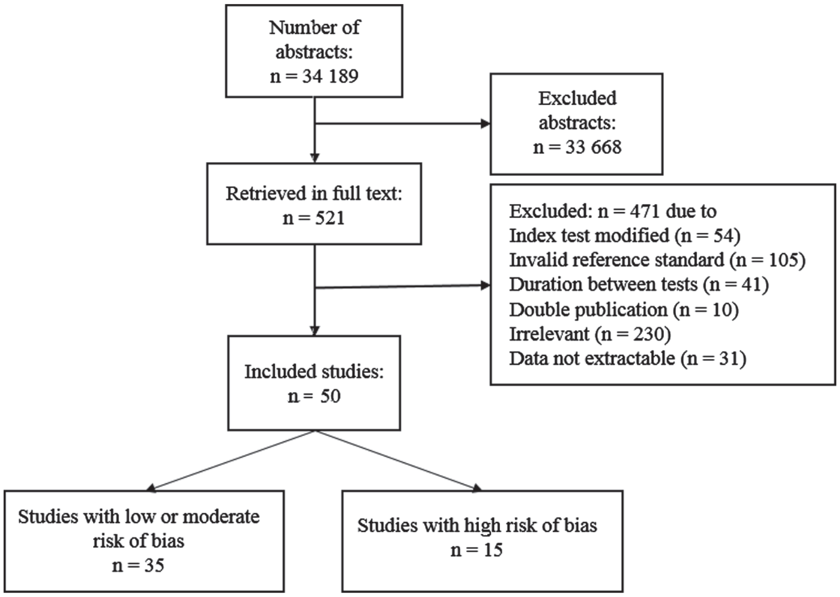
Number of abstracts: n = 34 189
Excluded abstracts: n = 33 668
Retrieved in full text: n = 521
Excluded: n = 471 due to
Index test modified (n = 54)
Invalid reference standard (n = 105)
Duration between tests (n = 41)
Double publication (n = 10)
Irrelevant (n = 230)
Data not extractable (n = 31)
Included studies: n = 50
Studies with low or moderate risk of bias n = 35
Studies with high risk of bias n = 15

Fig. 1. Flow chart of the literature selection.

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.

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

PCC, primary care centres.

Table 3. GRADE: Summary of findings for diagnostic accuracy of structured interviews with structured interviews as reference.

Instrument	Outcome	Studies (*n*); patients (*n*)	Average (95% CI)	Confidence in estimate	Reasons for downgrading confidence
SCID-I	Sensitivity	2; 256	86% (73–94%)	Low	Imprecision: − 2[*]
	Specificity	2; 256	92% (88–95%)	High	
MINI	Sensitivity	2; 663	95% (93–97%)	High	
	Specificity	2; 663	84% (80–87%)	High	
PRIME-MD	Sensitivity	2; 757	< 70%	Moderate	Inconsistency: − 1[†]
	Specificity	2; 757	85% (82–88%)	High	
CIDI	Sensitivity	1; 54	67%	Very low	Indirectness: − 2[‡]; Imprecision: − 1
	Specificity	1; 54	84%	Very low	Indirectness: − 2[‡]; Imprecision: − 1[§]
DIS	Sensitivity	1; 120	25%	Very low	Risk of bias: − 1; Indirectness: − [1‡]; Imprecision: − 1[§]
	Specificity	1; 120	90%	Very low	Risk of bias: − 1; Indirectness: − 1[‡];Imprecision: − 1[§]

[*] Lower 95% CI below benchmark.
[†] Reported sensitivities of 30% and 68%.
[‡] Few patients with narrow spectrum.
[§] Only one study.

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.

Study	Patients	Index test	Reference test	Sensitivity and specificity	Risk of bias
Bunevicius, 2012 (45)	*n* = 522 (cardiac rehabilitation clinic)	BDI-II	MINI	Sensitivity: 89%; specificity: 74%	Low
Dutton, 2004 (46)	*n* = 223 (PCC waiting room)	BDI-II	PRIME-MD	Sensitivity: 88%; specificity: 84%	Moderate
Poole, 2009 (43)	*n* = 36 (pain specialist centre)	BDI-II	SCID-I	Sensitivity: 100%; specificity: 50%	Moderate
Warmenhoven, 2011 (44)	*n* = 61 (oncology clinic)	BDI-II	PRIME-MD	Sensitivity: 90%; specificity: 64%	Moderate
Hall, 1999 (57)	*n* = 269 (breast cancer outpatients)	HADS	PSE	Sensitivity: 33%; specificity: 93%	Moderate
Haddad, 2013 (62)	*n* = 730 (cardiac patients, PCCs)	HADS	CIS-R	Sensitivity: 53%; specificity: 91%	Moderate
Lowe, 2004 (61)	*n* = 2050 (PCCs and hospital outpatients)	HADS	SCID-I	Sensitivity: 88%; specificity: 69%	Moderate
Orive, 2010 (58)	*n* = 167 (hospital waiting rooms)	HADS	PRIME-MD	Sensitivity: 86%; specificity 75%	Moderate
Silverstone, 1994 (59)	*n* = 189 (7 days after acute hospitalization)	HADS	SCAN	Sensitivity: 100%; specificity: 73%	Moderate
Stafford, 2007 (64)	*n* = 528 (cardiac outpatients)	HADS	MINI	Sensitivity: 46%; specificity: 92%	Moderate
Sultan, 2010 (99)	*n* = 370 (diabetes clinic)	HADS	MINI	Sensitivity: 53%; specificity: 86%	Moderate
Terluin, 2009 (60)	*n* = 370 (on sick leave, PCCs)	HADS	CIDI	Sensitivity: 93%; specificity: 39%	Low
Whelan Goodinson, 2009 (56)	*n* = 157 (hospital after traumatic brain injury)	HADS	SCID-I	Sensitivity: 62%; specificity: 92%	Moderate
Zoger, 2004 (63)	*n* = 98 (tinnitus outpatients)	HADS	SCID-I	Sensitivity: 80%; specificity: 94%	Moderate
Bombardier, 2012 (75)	*n* = 142 (spinal cord injuries)	PHQ-9	SCID-I	*Cut-off score of 10:* Sensitivity: 100%; specificity: 80%	Low
Cassin, 2013 (73)	*n* = 244 (bariatric surgery candidates)	PHQ-9	MINI	*Cut-off score of 10:*Sensitivity: 75%; specificity: 63%	Moderate
Cassin, 2013 (73)	*n* = 275 (bariatric surgery candidates)	PHQ-9	MINI	*Cut-off score of 10:*Sensitivity: 80%; specificity: 46%	Moderate
Diez-Quevedo (71)	*n* = 1003 (medical and surgical inpatients)	PHQ-9	SCID-I	*Algorithm:* Sensitivity: 84%; specificity: 92%	Moderate
Elderon, 2011 (74)	*n* = 1024 (cardiac outpatients)	PHQ-9	C-DIS	*Cut-off score of 10:* Sensitivity: 54%; specificity: 90%	Moderate
Haddad, 2013 (62)	*n* = 730 (cardiac outpatients, PCCs)	PHQ-9	CIS-R	*Algorithm:* Sensitivity: 59%; specificity: 95%. *Cut-off score of 10:* Sensitivity: 84%; specificity: 90%	Moderate
Henkel, 2004 (66)	*n* = 470 (PCCs)	PHQ-9	CIDI	*Algorithm:* Sensitivity: 78%; specificity: 85%	Moderate
Lowe, 2004 (61)	*n* = 2050 (outpatients medical hospital and PCCs)	PHQ-9	SCID-I	*Algorithm:* Sensitivity: 83%; specificity: 90%. *Cut-off score of 11:* Sensitivity: 90%; specificity: 77%	Moderate
Navines, 2012 (69)	*n* = 500 (HCV outpatients)	PHQ-9	SCID-I	*Algorithm:* Sensitivity: 72%; specificity: 99%	Low
Orive, 2010 (58)	*n* = 53 (hospital waiting rooms)	PHQ-9	PRIME-MD	*Cut-off score of 10:* Sensitivity: 68%; specificity: 89%	Moderate
Persoons, 2003 (68)	*n* = 97 (otolaryngology outpatients)	PHQ-9	MINI	*Algorithm:* Sensitivity: 68.8%; specificity: 94.4%	Moderate
Picardi, 2005 (70)	*n* = 141 (dermatology, inpatients)	PHQ-9	SCID-I	*Algorithm:* Sensitivity: 55%; specificity: 91%	Moderate
Spitzer, 1999 (42), Kroenke, 2001 (72)	*n* = 3000 (PCCs)	PHQ-9	SCID-I	*Algorithm:* Sensitivity: 73%; specificity: 98%. *Cut-off score of 10:* Sensitivity: 88%; specificity: 88%	Moderate
Stafford, 2007 (64)	*n* = 528 (cardiac outpatients)	PHQ-9	MINI	*Algorithm:* Sensitivity: 34%; specificity: 97%. *Cut-off score of 10:* Sensitivity: 54%; specificity: 91%	Moderate
Thekkumpurath, 2011 (67)	*n* = 782 (regional oncology clinics)	PHQ-9	SCID-I	*Algorithm:* Sensitivity: 56%; specificity: 96%. *Cut-off score of 10:* Sensitivity: 73%; specificity: 88%	Moderate

(Continued)

Table 4. (*Continued*).

Study	Patients	Index test	Reference test	Sensitivity and specificity	Risk of bias
Wittkampf, 2009 (65)	*n* = 689 (PCCs)	PHQ-9	SCID-I	*Algorithm:* Sensitivity: 68%; specificity: 95%. *Cut-off score of 10:* Sensitivity: 100%; specificity: 45%	Low
Caracciolo, 2002 (53)	*n* = 151 (2/3 orthopaedic disorders, 1/3 neurologic)	CES-D	SCID-I	Sensitivity: 100%; specificity: 57% (orthopaedic) and 36% (neurologic)	Moderate
Breslau, 1985 (54)	*n* = 308 (mothers to severely disabled children)	CES-D	DIS	Sensitivity: 88%; specificity: 73%	Moderate

MDD, major depression disorder; SCID-I, Structured Clinical Interview for DSM-IV-Axis I disorders; PCC, primary care centre; LEAD, Longitudinal Experts All Data; SCAN, Schedules for Clinical Assessment in Neuropsychiatry; CIDI, Composite International Diagnostic Interview; DIS, Diagnostic Interview Schedule; MINI, Mini International Neuropsychiatric Interview; PRIME-MD, Primary Care Evaluation of Mental Disorders; CIS-R, Revised Clinical Interview Schedule; BDI-II, Beck Depression Inventory II; HADS, Hospital Anxiety Depression Scale; PHQ-9, Patient Health Questionnaire – 9 items; and CES-D, Center for Epidemiological Studies-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.

Instrument	Outcome	Studies (*n*); patients (*n*)	Average (95% CI)	Confidence in estimate	Reasons for downgrading confidence
BDI-II, cut-off score of 14	Sensitivity	4; 824	92% (83–97%)	Moderate	Imprecision: − 1*
	Specificity	4; 824	72% (58–82%)	Very low	Inconsistency: − 2†;Imprecision: − 1*
HADS, cut-off score of 7	Sensitivity	10; 4928	70% (55–82%)	Low	Inconsistency: − 1*; Imprecision: − 1‡
	Specificity	10; 4928	83% (73–90%)	Low	Inconsistency: − 1*; Imprecision: − 1†
PHQ-9, algorithm-based	Sensitivity	11; 9990	69% (60–76%)	Moderate	Imprecision: − 1‡
	Specificity	11; 9990	95% (92–97%)	High	
PHQ-9, cut-off score of 10	Sensitivity	10; 9517	88% (77–94%)	Moderate	Inconsistency: − 1§
	Specificity	10; 9517	78% (65–88%)	Moderate	Imprecision: − 1‡
CES-D, cut-off score of 16	Sensitivity	2; 459	95% (83–99%)	Very low	Risk of bias: − 1‖; Indirectness: − 2¶
	Specificity	2; 459	33–73%	Very low	Risk of bias: − 1‖; Indirectness: − 2¶

*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.

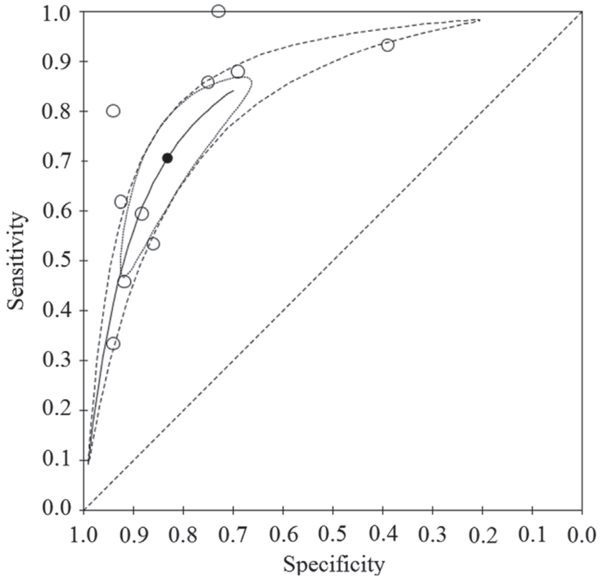

Fig. 2. Hierarchical summary receiver operating characteristic plot for the case finding of depression in adults using the Hospital Anxiety and Depression Scale (HADS) with a cut-off score of 7 with the diagnosis of depression by structured interview as the reference standard. Included studies are represented by white dots and the average estimate is represented by the black dot. The average sensitivity was 70% and the average specificity was 83%. The 95% confidence region is marked with a solid line and the prediction region is marked with a dotted line.

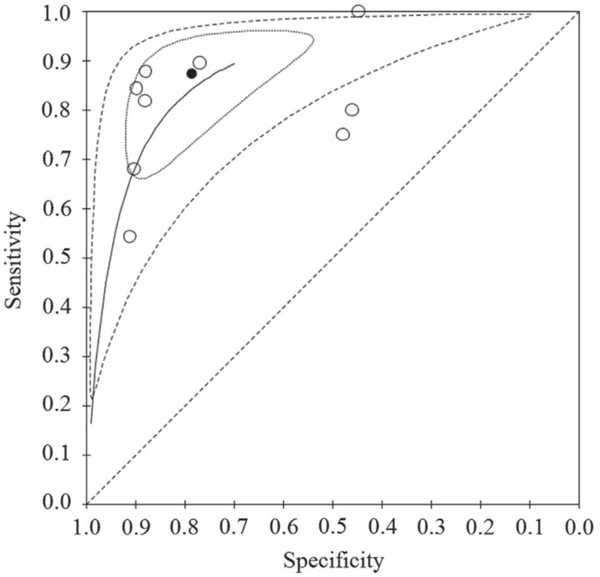

Fig. 3. Hierarchical summary receiver operating characteristic plot for the case finding of depression in adults using the Patient Health Questionnaire-9 (PHQ-9) with a cut-off score of 10 with the diagnosis of depression by structured interview as the reference standard. Included studies are represented by white dots and the average estimate is represented by the black dot. The average sensitivity was 88% and the average specificity was 78%. The 95% confidence region is marked with a solid line and the prediction region is marked with a dotted line.

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.

Instrument	Outcome	Studies (*n*); patients (*n*)	Summary estimate	Confidence in estimate	Reasons for grading down confidence
HDRS-17	Correlation with CGI-S	1 (98)	Spearman rank correlatio $n = 0.85$	Very low	Indirectness: -2*; Imprecision: -1†

*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 Brolund, 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

1. Smolders M, Laurant M, Verhaak P, Prins M, van Marwijk H, Penninx B, et al. Adherence to evidence-based guidelines for depression and anxiety disorders is associated with recording of the diagnosis. Gen Hosp Psychiatry 2009;31:460–9.
2. Spitzer RL. Psychiatric diagnosis: Are clinicians still necessary? Compr Psychiatry 1983;24:399–411.
3. Leckman JF, Sholomskas D, Thompson WD, Belanger A, Weissman MM. Best estimate of lifetime psychiatric diagnosis: A methodological study. Arch Gen Psychiatry 1982;39:879–83.
4. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: A meta-analysis. Lancet 2009;374:609–19.
5. Ramirez Basco M, Bostic JQ, Davies D, Rush AJ, Witte B, Hendrickse W, et al. Methods to improve diagnostic accuracy in a community mental health setting. Am J Psychiatry 2000;157:1599–605.
6. Miller PR, Dasher R, Collins R, Griffiths P, Brown F. Inpatient diagnostic assessments: 1. Accuracy of structured vs. unstructured interviews. Psychiatry Res 2001;105:255–64.
7. Taiminen T, Ranta K, Karlsson H, Lauerma H, Leinonen KM, Wallenius E, et al. Comparison of clinical and best-estimate

research DSM-IV diagnoses in a Finnish sample of first-admission psychosis and severe affective disorder. Nord J Psychiatry 2001;55:107–11.
8. Gilbody S, Sheldon T, House A. Screening and case-finding instruments for depression: A meta-analysis. CMAJ 2008;178:997–1003.
9. O'Connor EA, Whitlock EP, Beil TL, Gaynes BN. Screening for depression in adult patients in primary care settings: A systematic evidence review. Ann Internal Med 2009;151:793–803.
10. Knaup C, Koesters M, Schoefer D, Becker T, Puschner B. Effect of feedback of treatment outcome in specialist mental healthcare: Meta-analysis. Br J Psychiatry 2009;195:15–22.
11. Shaw EJ, Sutcliffe D, Lacey T, Stokes T. Assessing depression severity using the UK Quality and Outcomes Framework depression indicators: A systematic review. Br J Gen Pract 2013;63: e309–17.
12. Gilbody S, Sheldon T, Wessely S. Should we screen for depression? BMJ 2006;332:1027–30.
13. Brennan C, Worrall-Davies A, McMillan D, Gilbody S, House A. The Hospital Anxiety and Depression Scale: A diagnostic meta-analysis of case-finding ability. J Psychosom Res 2010;69: 371–8.
14. Meader N, Mitchell AJ, Chew-Graham C, Goldberg D, Rizzo M, Bird V, et al. Case identification of depression in patients with chronic physical health problems: A diagnostic accuracy meta-analysis of 113 studies. Br J Gen Pract 2011;61:e808–20.
15. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. CMAJ 2012;184:E191–6.
16. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469–76.
17. Nuevo R, Dunn G, Dowrick C, Vazquez-Barquero JL, Casey P, Dalgard OS, et al. Cross-cultural equivalence of the Beck Depression Inventory: A five-country analysis from the ODIN study. J Affect Disord 2009;114:156–62.
18. Lesser IM. Cultural considerations using the Structured Clinical Interview for DSM-III for mood and anxiety disorders assessment. J Psychopathol Behav Assess 1997;19:149–60.
19. Iwata N, Buka S. Race/ethnicity and depressive symptoms: A cross-cultural/ethnic comparison among university students in East Asia, North and South America. Soc Sci Med 2002;55: 2243–52.
20. SBU. Diagnostik och uppföljning av förstämningssyndrom: En systematisk litteraturöversikt. Report. Stockholm: Statens beredning för medicinsk utvärdering (SBU), 2012 212.
21. American Psychiatric Association (APA). Diagnostic and statistical manual of mental disorders. 4 ed. Washington: American Psychiatry Press; 1995.
22. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Ann Internal Med 2009;151:264–9, W64.
23. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol 2006;6:9.
24. Fleiss J. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–82.
25. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–58.
26. Macaskill P, Gatsonis CA, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10 Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis CA, editors. Cochrane handbook for systematic reviews of diagnostic test accuracy, Version 10: Cochrane Collaboration; 2010.
27. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865–84.
28. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-Disc: A software for meta-analysis of test accuracy data. BMC Med Res Methodol 2006;6:31.
29. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6.
30. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22–33;quiz 4–57.
31. Spitzer RL, Williams JB, Kroenke K, Linzer M, deGruy FV, 3rd, Hahn SR, et al. Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. JAMA 1994;272:1749–56.
32. Kessler RC, Ustun TB. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Methods Psychiatr Res 2004;13:93–121.
33. Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. Arch Gen Psychiatry 1981;38:381–9.
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.
36. Dahl AA, Kruger MB, Dahl NH, Karlson H, Knorring LV, Stordal E. SPIFA-A presentation of the Structured Psychiatric Interview for General Practice. Nord J Psychiatry 2009:1–11.
37. Addington D, Addington J, Maticka-Tyndale E, Joyce J. Reliability and validity of a depression rating scale for schizophrenics. Schizophr Res 1992;6:201–8.
38. Bech P, Rasmussen NA, Olsen LR, Noerholm V, Abildgaard W. The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. J Affect Disord 2001;66:159–64.
39. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–71.
40. Radloff LS. The CES-D Scale: A self-report depression scale for research in the general population. Appl Psychol Meas 1977;1: 385–401.
41. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361–70.
42. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA 1999;282:1737–44.
43. Poole H, White S, Blake C, Murphy P, Bramwell R. Depression in chronic pain patients: Prevalence and measurement. Pain Pract 2009;9:173–80.
44. Warmenhoven F, van Rijswijk E, Engels Y, Kan C, Prins J, van Weel C, et al. The Beck Depression Inventory (BDI-II) and a single screening question as screening tools for depressive disorder in Dutch advanced cancer patients. Support Care Cancer 2011;20: 319–24.
45. Bunevicius A, Staniute M, Brozaitiene J, Bunevicius R. Diagnostic accuracy of self-rating scales for screening of depression in coronary artery disease patients. J Psychosom Res 2012;72:22–5.
46. Dutton GR, Grothe KB, Jones GN, Whitehead D, Kendra K, Brantley PJ. Use of the Beck Depression Inventory-II with African American primary care patients. Gen Hosp Psychiatry 2004;26: 437–42.
47. Homaifar BY, Brenner LA, Gutierrez PM, Harwood JF, Thompson C, Filley CM, et al. Sensitivity and specificity of the Beck Depression Inventory-II in persons with traumatic brain injury. Arch Phys Med Rehabil 2009;90:652–6.
48. Di Benedetto M, Lindner H, Hare DL, Kent S. Depression following acute coronary syndromes: A comparison between the Cardiac Depression Scale and the Beck Depression Inventory II. J Psychosom Res 2006;60:13–20.
49. Ailey SH. The sensitivity and specificity of depression screening tools among adults with intellectual disabilities. J Mental Health Res Intellect Disab 2009;2:45–64.
50. Huffman JC, Doughty CT, Januzzi JL, Pirl WF, Smith FA, Fricchione GL. Screening for major depression in post-myocardial inf-

arction patients: Operating characteristics of the Beck Depression Inventory-II. Int J Psychiatry Med 2010;40:187–97.

51. Hopko DR, Bell JL, Armento ME, Robertson SM, Hunt MK, Wolf NJ, et al. The phenomenology and screening of clinical depression in cancer patients. J Psychosoc Oncol 2008;26:31–51.
52. De Souza J, Jones LA, Rickards H. Validation of self-report depression rating scales in Huntington's disease. Mov Disord 2009.
53. Caracciolo B, Giaquinto S. Criterion validity of the center for epidemiological studies depression (CES-D) scale in a sample of rehabilitation inpatients. J Rehabil Med 2002;34:221–5.
54. Breslau N. Depressive symptoms, major depression, and generalized anxiety: A comparison of self-reports on CES-D and results from diagnostic interviews. Psychiatry Res 1985;15:219–29.
55. Sultan S, Luminet O, Hartemann A. Cognitive and anxiety symptoms in screening for clinical depression in diabetes A systematic examination of diagnostic performances of the HADS and BDI-SF. J Affect Disord 2009.
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;114:94–102.
57. Hall A, A'Hern R, Fallowfield L. Are we using appropriate self-report questionnaires for detecting anxiety and depression in women with early breast cancer? Eur J Cancer 1999;35:79–85.
58. Orive M, Padierna JA, Quintana JM, Las-Hayas C, Vrotsou K, Aguirre U. Detecting depression in medically ill patients: Comparative accuracy of four screening questionnaires and physicians' diagnoses in Spanish population. J Psychosom Res 2010;69:399–406.
59. Silverstone PH. Poor efficacy of the Hospital Anxiety and Depression Scale in the diagnosis of major depressive disorder in both medical and psychiatric patients. J Psychosom Res 1994;38:441–50.
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.
61. Lowe B, Spitzer RL, Grafe K, Kroenke K, Quenter A, Zipfel S, et al. Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. J Affect Disord 2004;78:131–40.
62. Haddad M, Walters P, Phillips R, Tsakok J, Williams P, Mann A, et al. Detecting depression in patients with coronary heart disease: A diagnostic evaluation of the PHQ-9 and HADS-D in primary care, findings from the UPBEAT-UK study. PloS one 2013;8:e78493.
63. Zoger S, Svedlund J, Holgers KM. The Hospital Anxiety and Depression Scale (HAD) as a screening instrument in tinnitus evaluation. Int J Audiol 2004;43:458–64.
64. Stafford L, Berk M, Jackson HJ. Validity of the Hospital Anxiety and Depression Scale and Patient Health Questionnaire-9 to screen for depression in patients with coronary artery disease. Gen Hosp Psychiatry 2007;29:417–24.
65. Wittkampf K, van Ravesteijn H, Baas K, van de Hoogen H, Schene A, Bindels P, et al. The accuracy of Patient Health Questionnaire-9 in detecting depression and measuring depression severity in high-risk groups in primary care. Gen Hosp Psychiatry 2009;31:451–9.
66. Henkel V, Mergl R, Kohnen R, Allgaier AK, Moller HJ, Hegerl U. Use of brief depression screening tools in primary care: Consideration of heterogeneity in performance in different patient groups. Gen Hosp Psychiatry 2004;26:190–8.
67. Thekkumpurath P, Walker J, Butcher I, Hodges L, Kleiboer A, O'Connor M, et al. Screening for major depression in cancer outpatients: The diagnostic accuracy of the 9-item Patient Health Questionnaire. Cancer 2011;117:218–27.
68. Persoons P, Luyckx K, Desloovere C, Vandenberghe J, Fischler B. Anxiety and mood disorders in otorhinolaryngology outpatients presenting with dizziness: Validation of the self-administered PRIME-MD Patient Health Questionnaire and epidemiology. Gen Hosp Psychiatry 2003;25:316–23.
69. Navines R, Castellvi P, Moreno-Espana J, Gimenez D, Udina M, Canizares S, et al. Depressive and anxiety disorders in chronic hepatitis C patients: Reliability and validity of the Patient Health Questionnaire. J Affect Disord 2012;138:343–51.
70. Picardi A, Adler DA, Abeni D, Chang H, Pasquini P, Rogers WH, et al. Screening for depressive disorders in patients with skin diseases: A comparison of three screeners. Acta Derm Venereol 2005;85:414–9.
71. Diez-Quevedo C, Rangil T, Sanchez-Planell L, Kroenke K, Spitzer RL. Validation and utility of the Patient Health Questionnaire in diagnosing mental disorders in 1003 general hospital Spanish inpatients. Psychosom Med 2001;63:679–86.
72. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: Validity of a brief depression severity measure. J Gen Intern Med 2001;16: 606–13.
73. Cassin S, Sockalingam S, Hawa R, Wnuk S, Royal S, Taube-Schiff M, et al. Psychometric properties of the Patient Health Questionnaire (PHQ-9) as a depression screening tool for bariatric surgery candidates. Psychosomatics 2013;54:352–8.
74. Elderon L, Smolderen KG, Na B, Whooley MA. Accuracy and prognostic value of American Heart Association: Recommended depression screening in patients with coronary heart disease: Data from the Heart and Soul Study. Circulation Cardiovascular quality and outcomes 2011;4:533–40.
75. Bombardier CH, Kalpakjian CZ, Graves DE, Dyer JR, Tate DG, Fann JR. Validity of the Patient Health Questionnaire-9 in assessing major depressive disorder during inpatient spinal cord injury rehabilitation. Arch Phys Med Rehabil 2012;93: 1838–45.
76. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56–62.
77. Rush AJ, Giles DE, Schlesser MA, Fulton CL, Weissenburger J, Burns C. The Inventory for Depressive Symptomatology (IDS): Preliminary findings. Psychiatry Res 1986;18:65–87.
78. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979;134:382–9.
79. Zung WW. A self-rating depression scale. Arch Gen Psychiatry 1965;12:63–70.
80. Bech P, Kastrup M, Rafaelsen OJ. Mini-compendium of rating scales for states of anxiety depression mania schizophrenia with corresponding DSM-III syndromes. Acta Psychiatr Scand Suppl 1986;326:1–37.
81. Korner A, Lauritzen L, Abelskov K, Gulmann NC, Brodersen AM, Wedervang-Jensen T, et al. Rating scales for depression in the elderly: External and internal validity. J Clin Psychiatry 2007;68:384–9.
82. Rettew DC, Lynch AD, Achenbach TM, Dumenci L, Ivanova MY. Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. Int J Methods Psychiatr Res 2009;18:169–84.
83. Amorim P, Lecrubier Y, Weiller E, Hergueta T, Sheehan D. DSM-IH-R Psychotic Disorders: Procedural validity of the Mini International Neuropsychiatric Interview (MINI). Concordance and causes for discordance with the CIDI. Eur Psychiatry 1998;13:26–34.
84. Jones JE, Hermann BP, Barry JJ, Gilliam F, Kanner AM, Meador KJ. Clinical assessment of Axis I psychiatric morbidity in chronic epilepsy: A multicenter investigation. J Neuropsychiatry Clin Neurosci 2005;17:172–9.
85. Spitzer RL, Williams JB, Kroenke K, Hornyak R, McMurray J. Validity and utility of the PRIME-MD Patient Health Questionnaire in assessment of 3000 obstetric-gynecologic patients: The PRIME-MD Patient Health Questionnaire Obstetrics–Gynecology Study. Am J Obstet Gynecol 2000;183:759–69.
86. Maters GA, Sanderman R, Kim AY, Coyne JC. Problems in cross-cultural use of the Hospital Anxiety and Depression Scale: "No butterflies in the desert". PloS one 2013;8:e70975.
87. Kendel F, Wirtz M, Dunkel A, Lehmkuhl E, Hetzer R, Regitz-Zagrosek V. Screening for depression: Rasch analysis of the dimensional structure of the PHQ-9 and the HADS-D. J Affect Disord 2010;122:241–6.
88. Cameron IM, Crawford JR, Lawton K, Reid IC. Differential item functioning of the HADS and PHQ-9: An investigation of age, gender and educational background in a clinical UK primary care sample. J Affect Disord 2013;147:262–8.

89. Forkmann T, Gauggel S, Spangenberg L, Brahler E, Glaesmer H. Dimensional assessment of depressive severity in the elderly general population: Psychometric evaluation of the PHQ-9 using Rasch Analysis. J Affect Disord 2013;148:323–30.
90. Wang YP, Gorenstein C. Psychometric properties of the Beck Depression Inventory-II: A comprehensive review. Rev Bras Psiquiatr 2013;35:416–31.
91. Baas KD, Cramer AO, Koeter MW, van de Lisdonk EH, van Weert HC, Schene AH. Measurement invariance with respect to ethnicity of the Patient Health Questionnaire-9 (PHQ-9). J Affect Disord 2011;129:229–35.
92. Lecrubier Y, Sheehan DV, Weiller E, Amorim P, Bonora I, Harnett Sheehan K. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. Eur Psychiatry 1997;12: 224–31.
93. Sheehan D, Lecrubier Y, Harnett Sheehan K, Janavs J, Weiller E, Keskiner A. The validity of the Mini International Neuropsychitric Interview (MINI) according to the SCID-P and its reliability. Eur Psychiatry 1997;12:232–41.
94. Leopold KA, Ahles TA, Walch S, Amdur RJ, Mott LA, Wiegand-Packard L. Prevalence of mood disorders and utility of the PRIME-MD in patients undergoing radiation therapy. Radiat Oncol Biol Phys 1998;42:1105–12.
95. Loerch B, Szegedi A, Kohnen R, Benkert O. The primary care evaluation of mental disorders (PRIME-MD), German version: a comparison with the CIDI. J Psychiatr Res 2000;34:211–20.
96. Booth BM, Kirchner JE, Hamilton G, Harrell R, Smith GR. Diagnosing depression in the medically ill: validity of a lay-administered structured diagnostic interview. J Psychiatr Res 1998;32:353–60.
97. Hasin DS, Grant BF. Diagnosing depressive disorders in patients with alcohol and drug problems: a comparison of the SADS-L and the DIS. J Psychiatr Res 1987;21:301–11.

Agneta Pettersson, M.Sc., Department of Learning, Informatics, Medical Education and Ethics, Karolinska Institutet, and Swedish Council on Health Technology Assessment, Stockholm, Sweden.

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.

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