Accuracy of the Hospital Anxiety and Depression Scale Depression subscale (HADS-D) to screen for major depression: systematic review and individual participant data meta-analysis

Yin Wu,[1,2] Brooke Levis,[3] Ying Sun,[1] Chen He,[1] Ankur Krishnan,[1] Dipika Neupane,[1] Parash Mani Bhandari,[1] Zelalem Negeri,[1] Andrea Benedetti,[4-6] Brett D Thombs,[1,2,4,6-9] on behalf of the DEPRESsion Screening Data (DEPRESSD) HADS Group

For numbered affiliations see end of the article.

Correspondence to: B D Thombs brett.thombs@mcgill.ca (ORCID 0000-0002-5644-8432)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2021;373:n972 http://dx.doi.org/10.1136/bmj.n972

Accepted: 9 April 2021

ABSTRACT

OBJECTIVE

To evaluate the accuracy of the depression subscale of the Hospital Anxiety and Depression Scale (HADS-D) to screen for major depression among people with physical health problems.

DESIGN

Systematic review and individual participant data meta-analysis.

DATA SOURCES

Medline, Medline In-Process and Other Non-Indexed Citations, PsycInfo, and Web of Science (from inception to 25 October 2018).

REVIEW METHODS

Eligible datasets included HADS-D scores and major depression status based on a validated diagnostic interview. Primary study data and study level data extracted from primary reports were combined. For HADS-D cut-off thresholds of 5-15, a bivariate random effects meta-analysis was used to estimate pooled sensitivity and specificity, separately, in studies that used semi-structured diagnostic interviews (eg, Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders*), fully structured interviews (eg, Composite International Diagnostic Interview), and the Mini International Neuropsychiatric Interview. One stage meta-regression was used to examine whether accuracy was associated with reference standard categories and the characteristics of participants. Sensitivity analyses were done to assess whether including published results from studies that did not provide raw data influenced the results.

RESULTS

Individual participant data were obtained from 101 of 168 eligible studies (60%; 25 574 participants (72% of eligible participants), 2549 with major depression). Combined sensitivity and specificity was maximised at a cut-off value of seven or higher for semi-structured interviews, fully structured interviews, and the Mini International Neuropsychiatric Interview. Among studies with a semi-structured interview (57 studies, 10 664 participants, 1048 with major depression), sensitivity and specificity were 0.82 (95% confidence interval 0.76 to 0.87) and 0.78 (0.74 to 0.81) for a cut-off value of seven or higher, 0.74 (0.68 to 0.79) and 0.84 (0.81 to 0.87) for a cut-off value of eight or higher, and 0.44 (0.38 to 0.51) and 0.95 (0.93 to 0.96) for a cut-off value of 11 or higher. Accuracy was similar across reference standards and subgroups and when published results from studies that did not contribute data were included.

CONCLUSIONS

When screening for major depression, a HADS-D cut-off value of seven or higher maximised combined sensitivity and specificity. A cut-off value of eight or higher generated similar combined sensitivity and specificity but was less sensitive and more specific. To identify medically ill patients with depression with the HADS-D, lower cut-off values could be used to avoid false negatives and higher cut-off values to reduce false positives and identify people with higher symptom levels.

TRIAL REGISTRATION

PROSPERO CRD42015016761.

WHAT IS ALREADY KNOWN ON THIS TOPIC

The Hospital Anxiety and Depression Scale Depression subscale (HADS-D) is the most commonly used screening tool for depression in medically ill patients, with cut-off values of eight or higher or 11 or higher used as standards to identify possible or probable depression

The only previous meta-analysis on the accuracy of HADS for detecting major depression in all populations included 11 studies (1735 participants) up to 2006

WHAT THIS STUDY ADDS

At a HADS-D cut-off value of seven or higher, combined sensitivity and specificity were maximised (82%, 78%), based on 101 studies; sensitivity and specificity were 74% and 84% for a HADS-D cut-off value of eight or higher and 44% and 95% for a cut-off value of 11 or higher

Results did not differ across reference standards or participant characteristics, including age, sex, human development index levels, and recruitment setting of participants

A web based knowledge translation tool is available to estimate the expected number of positive screens, and true and false screening outcomes based on study results (depressionscreening100.com/hads-d)

Introduction

Major depressive disorder is present in 10-20% of patients with acute or chronic medical conditions and is associated with a poor prognosis.[1-6] Healthcare providers in non-psychiatric settings, where most of the care for depression is provided, might have relatively little formal mental health training.[7] Mental healthcare could be inconsistently delivered, particularly outside of primary care.[8 9] Many depressed patients are not identified, and a high proportion of patients treated for depression do not meet diagnostic criteria.[10-12]

Recognising depression can be particularly difficult in people with a physical illness, and some symptoms, such as fatigue, changes in appetite, and trouble sleeping, are common in both depression and many medical conditions.[7] Although controversial, screening for depression is sometimes used to identify people not previously recognised as having depression, including in chronic conditions and cancer.[13-19] Screening for depression involves giving short questionnaires to people not already known or suspected of having depression, with cut-off thresholds on the screening questionnaires to distinguish positive from negative screening results, and then assessing those with positive results further to determine whether the criteria for depression are met.[14-18]

The Hospital Anxiety and Depression Scale (HADS)[20] was developed to help identify anxiety disorders and depression in people with a physical illness. To avoid overlap with physical disorders, the HADS does not include somatic symptoms, such as insomnia, loss of appetite, or fatigue. The depression subscale of the HADS (HADS-D) is the most commonly used screening tool for depression in medically ill patients[21] and is one of several validated measures recommended for assessing the severity of depressive symptoms by the United Kingdom National Institute for Health and Care Excellence (NICE).[22] In the initial HADS-D validation study (100 participants, 12 with major depression),[20] which has been cited over 37 000 times since publication in 1983 (Google Scholar), the developers suggested that a cut-off value of eight or higher could be used to identify possible depression and a cut-off value of 11 or higher for probable depression. These cut-off values have since been used as standards in research and practice.[23-25]

Primary studies on the accuracy of HADS-D screening have been limited by samples too small to generate precise estimates; inability to conduct subgroup analyses; selective reporting of results from study specific "optimal cut-off values" that seem more accurate than standard cut-off values in a given sample; and including patients who would not be screened in practice because of a previous diagnosis or treatment for depression.[23-29] The only previous meta-analysis on the accuracy of the HADS-D for detecting major depression that was not restricted to subpopulations (eg, cancer, palliative care) included 11 eligible studies (1735 participants) up to June 2006.[23] Analyses of accuracy at cut-off values from eight or higher to 11 or higher were based on only six to seven studies, however, because not all of the 11 studies reported results for each cut-off value. Overall, 39% of otherwise eligible studies could not be included in any meta-analyses of the HADS-D or HADS for anxiety because results for commonly used cut-off values were not reported. Subgroup analyses were not possible because they were not available in primary studies. Also, results were combined across reference standard diagnostic interviews, even though important differences in design and structure exist. Recent studies have shown that different diagnostic interview formats have substantively different likelihoods of classifying major depression.[30-33]

An individual participant data meta-analysis (IPDMA)[34] involves a standard systematic review, followed by synthesis of line-by-line participant data from primary studies, rather than aggregated summary data. The advantages of an IPDMA with the HADS-D are the ability to include data from studies that administered the HADS-D and a diagnostic interview but did not publish accuracy results; to carry out subgroup analyses; to evaluate results excluding participants already diagnosed with or treated for depression who would not be screened in practice; and overcoming bias from selective cut-off reporting by including estimates of accuracy for all of the relevant cut-off values from the studies included in the meta-analysis. Our objectives were to evaluate the accuracy of the HADS-D to screen for depression, separately by different types of reference standards, prioritising semi-structured interviews; and to investigate whether accuracy differed according to age, sex, medical condition, country human development index, and recruitment setting or according to whether patients with previously diagnosed depression were included.

Methods

This IPDMA was registered in PROSPERO (CRD-42015016761), a protocol was published,[35] and the results were described according to the reporting guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of diagnostic test accuracy (PRISMA-DTA)[36] and PRISMA of individual participant data (PRISMA-IPD).[37] The methods were similar to our previously published IPDMAs of the accuracy of the Patient Health Questionaire-9[38] and the Edinburgh Postnatal Depression Scale.[39]

Dataset eligibility

Datasets from articles in any language were eligible if they included a diagnostic classification for current major depressive disorder or major depressive episode according to the criteria of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM)[40-43] or the International Classification of Diseases (ICD)[44] based on a validated semi-structured or fully structured interview; total scores for the HADS-D were included; the diagnostic interview and HADS-D were done within two weeks of each other, because the diagnostic criteria of the DSM and ICD for major depression specify that symptoms must have been present in the past two weeks; participants were aged 18 or older and were not recruited from youth or psychiatric settings; and participants were not recruited because they were identified as having symptoms of depression because screening is done to identify previously undiagnosed patients. Datasets where not all participants were eligible were included if the primary data allowed selection of eligible participants.

Search strategy and study selection

A medical librarian searched Medline, Medline In-Process and Other Non-Indexed Citations, and PsycInfo through OvidSP, and Web of Science through ISI Web of Knowledge, from inception to 25 October 2018 with a peer reviewed[45] search strategy (supplementary methods A). We also reviewed reference lists of relevant reviews and queried contributing authors about non-published studies. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA). After de-duplication, unique citations were uploaded into DistillerSR (Evidence Partners, Ottawa, ON, Canada) for tracking search results.

Pairs of investigators independently reviewed titles and abstracts for eligibility. If either believed that a study was potentially eligible, a full text review was done by pairs of investigators independently, with disagreements resolved by consensus, consulting a third investigator when necessary. Translators were consulted if team members were not fluent in the language of an article.

Data contribution, extraction, and synthesis

Authors of eligible datasets were invited to contribute de-identified primary data. We emailed corresponding authors of eligible primary studies at least three times, if necessary. If we did not receive a response, we emailed coauthors and attempted to contact corresponding authors by telephone.

Diagnostic interview and country were extracted from published reports by pairs of investigators independently, with disagreements resolved by consensus. Countries were categorised as "very high," "high", or "low-medium" development based on the United Nation's human development index for the country, for the year the study was published. The human development index is a statistical composite index that includes indicators of life expectancy, education, and income.[46] Participant level data extracted from the studies in the meta-analysis included age, sex, diagnosis of cancer, patient care setting, HADS-D scores, and major depression status (with or without major depression). We focused on cancer and not other medical conditions because not enough studies or data existed for analyses of other conditions. For defining major depression, we considered major depressive disorder or major depressive episode based on the DSM or ICD. If more than one was reported, we prioritised major depressive episode over major depressive disorder because screening would attempt to detect depressive episodes and further interview would determine if the episode was related to major depressive disorder, bipolar disorder, or persistent depressive disorder. We prioritised the DSM over the ICD.

Individual participant data were converted to a standard format and combined into one dataset with study level data. We compared the published characteristics of the participants and the estimates of the accuracy of the diagnoses with the results from the raw datasets and resolved any discrepancies by consulting the original investigators.

Risk of bias assessment

Risk of bias of the studies included in the meta-analysis was assessed by two investigators independently with the QUality Assessment of Diagnostic Accuracy Studies-2 tool (QUADAS-2; supplementary methods B).[47] Any discrepancies were resolved by consensus, and a third investigator was involved if necessary. Risk of bias was coded at both the study and participant levels because some data (eg, time between index test and reference standard) might have differed among participants from the same study.

Statistical analyses

We conducted three sets of analyses. Firstly, we separately pooled estimated sensitivity and specificity across HADS-D cut-off values for studies that used semi-structured reference standard interviews (Structured Clinical Interview for the DSM (SCID),[48] Schedules for Clinical Assessment in Neuropsychiatry,[49] Schedule for Affective Disorders and Schizophrenia,[50] Monash Interview for Liaison Psychiatry[51]), fully structured interviews (Composite International Diagnostic Interview (CIDI),[52] Diagnostic Interview Schedule[53]), and the Mini International Neuropsychiatric Interview (MINI).[54 55] The MINI was treated as a separate reference standard category throughout all analyses. We analysed studies that used different types of reference standards separately because they have different designs and performance characteristics. We found in previous IPDMAs that, compared with the semi-structured SCID interview, participants evaluated with the brief MINI were substantially more likely to be classified as having major depression. When the fully structured CIDI was used, participants with lower depressive symptom severity were more likely to be classified with major depression, but the opposite was true in those with greater symptom severity.[30-33] Semi-structured interviews should be carried out by experienced diagnosticians and are considered to most closely replicate clinical diagnostic procedures.[56-58]

For each reference standard category (semi-structured, fully structured, and the MINI), for HADS-D cut-off values of 5-15 separately, bivariate random effects models were fitted from Gauss-Hermite quadrature.[59] This two stage meta-analytic approach models sensitivity and specificity simultaneously and accounts for the correlation between them and the precision of estimates within studies. We constructed empirical receiver operating characteristic curves based on pooled sensitivity and specificity estimates, and calculated area under the curve values for each reference standard category. Also, we conducted one stage meta-regressions with interactions between reference standard category (reference category: semi-structured) and accuracy coefficients (logit(sensitivity) and logit(1–specificity)). To present positive and negative predictive values for the cut-off value that maximised combined sensitivity and specificity, and for standard cut-off values of eight or higher and 11 or higher, we generated nomograms for an assumed prevalence of major depression of 5-25% (based on

the prevalence of major depressive disorder in medical patients of 10-20% ±5%).[1-6]

To investigate heterogeneity, for each category of reference standard we generated forest plots of sensitivity and specificity for each study for the cut-off value that maximised combined sensitivity and specificity. Although no well established methods exist to quantify levels of heterogeneity in diagnostic test accuracy meta-analyses,[36 60] we quantified heterogeneity by reporting estimated variances of the random effects for sensitivity and specificity (τ^2) and by estimating R, the ratio of the estimated standard deviation of the pooled sensitivity (or specificity) from the random effects model to that from the corresponding fixed effects model.[61]

Secondly, to investigate whether the accuracy of HADS-D screening differs according to the characteristics of the participants, we conducted one stage meta-regressions separately by reference standard category (semi-structured, fully structured, and the MINI), where we interacted all subgroup variables (age (measured continuously), sex (reference category=female)), country human development index (reference category=very high), diagnosis of cancer (reference category=no), and recruitment setting for participants (reference category=inpatient specialty care) with logit(sensitivity) and logit(1–specificity). These models were restricted to the subset of studies that had complete data for all relevant variables. This method resulted in a loss of 520 (8%) participants that did not have data for age or sex from semi-structured interview studies, two participants (0.1%) from fully structured interview studies, and 88 participants (1%) from MINI studies. Also, for each reference standard, we estimated sensitivity and specificity in participants verified as not having been diagnosed or receiving treatment for mental health problems, and we compared the accuracy results with the results in all participants. This comparison was conducted because in practice, screening is done to identify previously undiagnosed people with major depression. Screening is never done in participants currently diagnosed or receiving treatment for mental health problems. The inclusion of patients who would not be screened in practice could bias the estimates of diagnostic accuracy, but many primary studies did not record the diagnostic or treatment status of the participants. Thus we evaluated whether results with all participants differed from results when only studies and participants where treatment and diagnostic status were known were analysed. Analytically, we calculated the confidence intervals of the differences by bootstrapping to account for the overlap in the two groups being compared.

For analysis of the possible influence of risk of bias, we added interactions of QUADAS-2 signalling item responses with logit(sensitivity) and logit(1–specificity) to the main one stage meta-regression models in each reference standard category separately. This method allowed us to compare the accuracy of the HADS-D by subgroups based on QUADAS-2 items for all items with at least 100 participants with major depression and 100 without major depression, categorised as having a low versus an unclear or high risk of bias.

Thirdly, in sensitivity analyses, we combined the accuracy results of the IPDMA with published results from studies that did not contribute individual participant data for each reference standard category, semi-structured, fully structured, and the MINI. Based on the publication of eligible accuracy results from studies that did not contribute data, we conducted this analysis for a HADS-D cut-off value of eight or higher in studies that used a semi-structured reference standard; for HADS-D cut-off values of eight or higher and 11 or higher in studies that used a fully structured reference standard; and for HADS-D cut-off values of eight or higher and 11 or higher in studies that used the MINI. All analyses were done in R (R version R 3.4.1[62] and R Studio version 1.0.143[63]) with the glmer function within the lme4 package.[64]

Patient and public involvement

No patients were involved in the initial development of the research question, outcome measures, or study design. Since study inception, Dr Sarah Markham has joined the DEPRESSD group as a patient collaborator. She provided comments on the draft manuscript. We have no plans to disseminate the results of the research to study participants or the relevant patient communities. An online knowledge translation tool, intended for clinicians (the end users of the HADS-D screening tool) and other interested parties, however, has been made available at depressionscreening100.com/hads-d. The tool allows clinicians to estimate the expected number of positive screens as well as true and false screening outcomes based on study results.

Results

Search results and dataset inclusion

We found 12 830 unique titles and abstracts from the database search. Of these, 12 300 were excluded after review of the title and abstract and 301 after a full text review (fig 1), resulting in 229 eligible articles from 158 unique participant samples. Of these, 92 (58%) contributed datasets (fig 1). Among 14 eligible studies published before 2000, only one (7%) contributed a dataset; in studies published between 2000 and 2009, 31 of 57 (54%) contributed datasets; and in studies published between 2010 and 2018, 60 of 87 (69%) contributed datasets. Supplementary table A provides the reasons for exclusion of the 301 articles after a full text review. Authors of the studies included in the meta-analysis contributed data from 10 other studies that the search did not retrieve, for a total of 102 datasets. Supplementary table B shows the characteristics of the primary studies that contributed data and the eligible studies that did not provide datasets. Of 31 535 participants in 168 eligible published studies, 22 600 (72%) were included. One dataset initially included was excluded from this study

because it had no participants with major depression and therefore could not be included in the bivariate random effects models. Thus a total of 101 datasets (22 574 participants, 2549 with major depression) were included in this study.

Of the 101 studies included, 57 used semi-structured reference standards, including 53 that used the SCID; 12 used fully structured reference standards (excluding the MINI), including 11 that used the CIDI; and 32 used the MINI (table 1 and table 2).

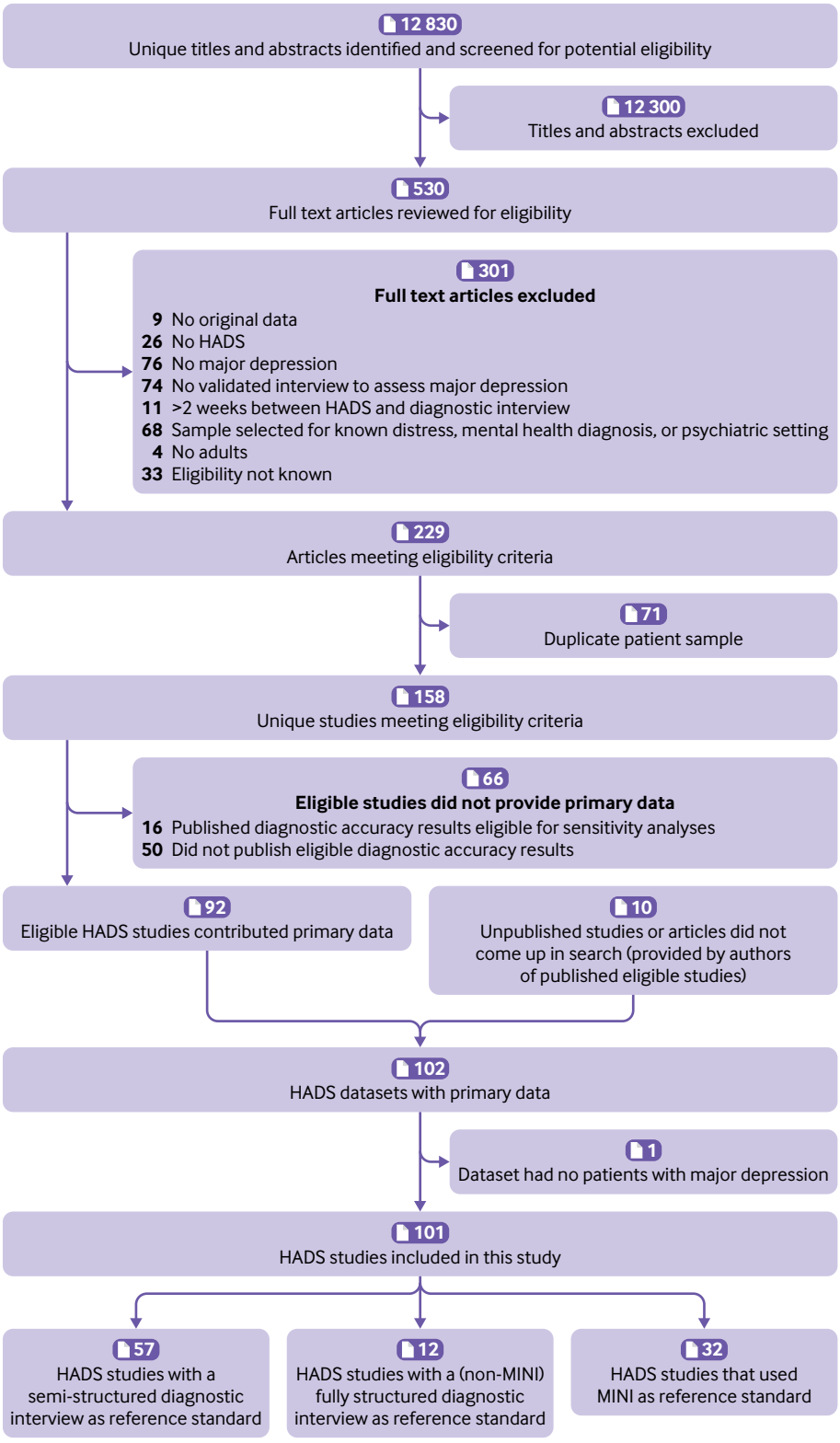

Fig 1 | Flow diagram of study selection process. HADS=Hospital Anxiety and Depression Scale; MINI=Mini International Neuropsychiatric Interview

Table 1 | Participant data by diagnostic interview

Diagnostic interview	No of studies	No of participants	Major depression (No (%))
Semi-structured			
SCID	53	10 029	983 (10)
SCAN	1	50	12 (24)
SADS	1	56	9 (16)
MILP	2	529	44 (8)
Fully structured			
CIDI	11	3705	327 (9)
DIS	1	194	11 (6)
MINI	32	8011	1163 (15)
Total	101	22 574	2549 (11)

CIDI=Composite International Diagnostic Interview; DIS= Diagnostic Interview Schedule; MILP= Monash Interview for Liaison Psychiatry; MINI=Mini International Neuropsychiatric Interview; SADS=Schedule for Affective Disorders and Schizophrenia; SCAN=Schedules for Clinical Assessment in Neuropsychiatry; SCID=Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders*.

HADS-D sensitivity and specificity by reference standard category

Table 3 shows the sensitivity and specificity estimates for cut-off values of 5-15 by reference standard category. Combined sensitivity and specificity was maximised at a cut-off value of seven or higher for semi-structured interviews, fully structured interviews, and the MINI. For semi-structured interviews, sensitivity and specificity were 0.82 (95% confidence interval 0.76 to 0.87) and 0.78 (0.74 to 0.81) for a cut-off value of seven or higher, 0.74 (0.68 to 0.79) and 0.84 (0.81 to 0.87) for a cut-off value of eight or higher, and 0.44 (0.38 to 0.51) and 0.95 (0.93 to 0.96) for a cut-off value of 11 or higher. Figure 2 shows receiver operating characteristic curves and area under the curve values. The area under the curve was 0.87 for semi-structured interviews, 0.85 for fully structured diagnostic interviews, and 0.83 for the MINI. We found no significant differences in accuracy by reference standard category that held across all cut-off values (supplementary table C).

Of the 66 published studies that did not contribute datasets, 16 published eligible accuracy results but only 12 published results for cut-off values of seven or higher, eight or higher, or 11 or higher (supplementary table B2; semi-structured interview=8, fully structured interview=2, MINI=2). Supplementary tables D1-D3 show that estimates were similar when these results were included.

Figure 3 shows nomograms of positive and negative predictive values for cut-off values of seven or higher, eight or higher, and 11 or higher for the semi-structured reference standard. For a prevalence of major depression of 5-25%, positive predictive values for a cut-off value of seven or higher compared with semi-structured interviews ranged from 17% to 56%, and negative predictive values ranged from 93% to 99%. Positive predictive values ranged from 20% to 61% for a cut-off value of eight or higher and from 32% to 75% for a cut-off value of 11 or higher; negative predictive values ranged from 91% to 98% for a cut-off value of eight or higher and from 84% to 97% for a cut-off value of 11 or higher. Ranges were similar for fully structured and MINI reference standard interviews.

Heterogeneity analyses suggested moderate heterogeneity across the studies. Supplementary figure A shows forest plots of sensitivity and specificity, and supplementary table E shows τ^2 and R values.

HADS-D accuracy among subgroups and by risk of bias

Sensitivity and specificity estimates were not significantly different for participants identified as not currently diagnosed or receiving mental health treatment compared with all participants across reference standard categories (supplementary table F). We found no significant differences in accuracy that were consistent across reference standard categories

Table 2 | Participant data by subgroup*

	Semi-structured diagnostic interviews			Fully structured diagnostic interviews			MINI		
Participant subgroup	No of studies	No of participants	Major depression (No (%))	No of studies	No of participants	Major depression (No (%))	No of studies	No of participants	Major depression (No (%))
All participants	57	10 664	1048 (10)	12	3899	338 (9)	32	8011	1163 (15)
Participants not currently diagnosed or receiving treatment for a mental health problem	23	3354	204 (6)	3	2069	123 (6)	13	2590	262 (10)
Age <60	52	5827	655 (11)	12	2274	257 (11)	31	4879	722 (15)
Age ≥60	52	4690	340 (7)	10	1623	81 (5)	32	3045	427 (14)
Women	55	6123	619 (10)	12	1807	179 (10)	32	3874	675 (17)
Men	52	4450	397 (9)	11	2092	159 (8)	29	4132	488 (12)
Very high country human development index	54	10 528	1013 (10)	8	3453	227 (7)	28	7684	1077 (14)
High country human development index	3	136	35 (26)	4	446	111 (25)	4	327	86 (26)
Participants diagnosed with cancer	17	3084	247 (8)	6	2058	123 (9)	4	663	67 (10)
Inpatient specialty care	35	6008	631 (11)	9	2040	145 (7)	12	5390	806 (15)
Outpatient specialty care	19	3650	466 (13)	5	1859	193 (10)	16	1826	305 (17)
Non-medical care	4	1225	82 (7)	—	—	—	3	695	35 (5)
Inpatient and outpatient mixed	2	320	34 (11)	—	—	—	1	100	17 (17)

MINI=Mini International Neuropsychiatric Interview.
*Some variables were coded at the study level and others at the participant level and so the number of studies does not always add up to the total number in the reference category.

Table 3 | Comparison of sensitivity and specificity estimates for each reference standard category

	Semi-structured reference standard*		Fully structured reference standard†		MINI reference standard‡	
Cut-off	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
5	0.92 (0.87 to 0.95)	0.61 (0.57 to 0.66)	0.89 (0.76 to 0.95)	0.57 (0.46 to 0.67)	0.88 (0.84 to 0.91)	0.63 (0.60 to 0.66)
6	0.88 (0.83 to 0.92)	0.70 (0.66 to 0.74)	0.86 (0.73 to 0.93)	0.66 (0.55 to 0.76)	0.82 (0.78 to 0.85)	0.72 (0.69 to 0.75)
7	0.82 (0.76 to 0.87)	0.78 (0.74 to 0.81)	0.81 (0.67 to 0.9)	0.73 (0.63 to 0.81)	0.75 (0.70 to 0.79)	0.80 (0.77 to 0.82)
8	0.74 (0.68 to 0.79)	0.84 (0.81 to 0.87)	0.70 (0.56 to 0.81)	0.83 (0.72 to 0.90)	0.66 (0.61 to 0.71)	0.86 (0.84 to 0.88)
9	0.64 (0.58 to 0.69)	0.88 (0.86 to 0.91)	0.59 (0.48 to 0.70)	0.88 (0.79 to 0.94)	0.55 (0.49 to 0.60)	0.91 (0.89 to 0.92)
10	0.55 (0.49 to 0.60)	0.92 (0.90 to 0.94)	0.48 (0.39 to 0.56)	0.92 (0.85 to 0.96)	0.44 (0.38 to 0.50)	0.93 (0.92 to 0.95)
11	0.44 (0.38 to 0.51)	0.95 (0.93 to 0.96)	0.37 (0.30 to 0.45)	0.94 (0.89 to 0.97)	0.34 (0.29 to 0.40)	0.96 (0.95 to 0.97)
12	0.35 (0.29 to 0.41)	0.96 (0.95 to 0.97)	0.25 (0.19 to 0.34)	0.96 (0.92 to 0.98)	0.28 (0.23 to 0.34)	0.97 (0.96 to 0.98)
13	0.27 (0.22 to 0.32)	0.98 (0.97 to 0.98)	0.19 (0.14 to 0.25)	0.97 (0.94 to 0.98)	0.19 (0.15 to 0.25)	0.98 (0.98 to 0.99)
14§	0.20 (0.16 to 0.25)	0.99 (0.98 to 0.99)	0.09 (0.04 to 0.19)	0.98 (0.96 to 0.99)	0.15 (0.15 to 0.15)	0.99 (0.99 to 0.99)
15	0.15 (0.11 to 0.19)	0.99 (0.99 to 1.00)	0.07 (0.03 to 0.14)	0.98 (0.97 to 0.99)	0.10 (0.08 to 0.14)	0.99 (0.99 to 1.00)

MINI=Mini International Neuropsychiatric Interview.
*Number of studies, participants, and participants with major depression are 57, 10 664, and 1048, respectively.
†Number of studies, participants, and participants with major depression are 12, 3899, and 338, respectively.
‡Number of studies, participants, and participants with major depression are 32, 8011, and 1163, respectively.
§Among studies with the MINI, the default optimiser in glmer failed at this cut-off value and bobyqa was used instead.

for any participant characteristic (supplementary table G).

Supplementary table H shows QUADAS-2 ratings for the studies included. No QUADAS-2 domain items were consistently associated with differences in estimates of sensitivity or specificity for the semi-structured, fully structured, and MINI reference standard categories (supplementary table G).

Discussion

Our main finding was that combined sensitivity (82%) and specificity (78%) was maximised at a HADS-D cut-off value of seven or higher among 57 studies that used semi-structured interviews, which are designed to be used by trained mental health professionals to replicate diagnostic procedures as closely as possible. At cut-off values of eight or higher and 11 or higher, which are often recommended for screening for depression,[22] sensitivity and specificity were 74% and 84%, and 44% and 95%, respectively.

For the cut-off values that we examined, HADS-D sensitivity was 1-11% higher compared with semi-structured interviews than fully structured interviews (excluding the MINI), and 4-11% higher when compared with the MINI. Specificity estimates were similar across different reference standards. We found no significant differences in accuracy between subgroups that replicated across reference standard categories, although some subgroups had limited numbers of participants and patients. The results did not differ when patients previously diagnosed as having depression or receiving treatment for a mental health problem were excluded.

In the only previous general HADS-D meta-analysis,[23] which aggregated published data from 11 studies and combined reference standards without adjustment, sensitivity and specificity were 0.82 (95% confidence interval 0.73 to 0.89) and 0.74 (0.60 to 0.84) for a cut-off value of eight or higher and 0.56 (0.40 to 0.71) and 0.92 (0.79 to 0.97) for a cut-off value of 11 or higher. The results were not reported for a cut-off value of seven or higher. Our results differed substantially. Of the 57 studies that used semi-structured interviews, pooled sensitivity for HADS-D was lower and specificity was higher for the recommended cut-off values of eight or higher and 11 or higher. Differences in results between our IPDMA and the previous meta-analysis might be because our meta-analysis included a much larger number (101 *v* 11) of primary studies, including 57 with a semi-structured reference standard, and incorporated data from all cut-off values for all of the studies included. In contrast, the previous general HADS-D aggregate data meta-analysis,[23] and two other meta-analyses that used subsets of studies of people with cancer or in palliative care,[24 25] combined reference standards without adjustment and only included studies that reported results on commonly used cut-off values of eight or higher and 11 or higher,[23] or fitted one bivariate accuracy model that included results from different optimal cut-off values from different primary studies.[24 25]

The finding that the combined sensitivity and specificity of the HADS-D was highest when compared with the semi-structured reference standard is consistent

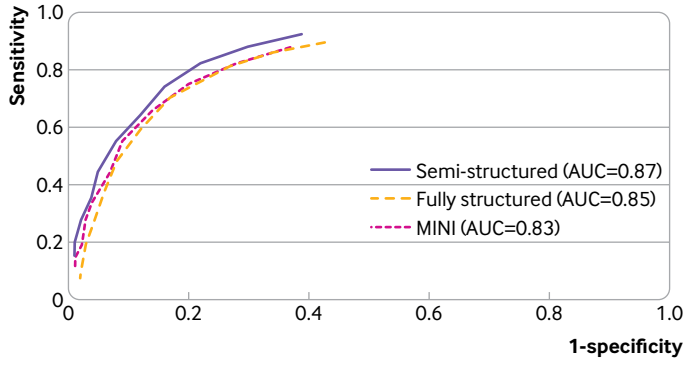

Fig 2 | Receiver operating characteristic (ROC) curves and area under the curve (AUC) values for each reference standard category. ROC curves comparing sensitivity and specificity estimates for the Hospital Anxiety and Depression Scale Depression subscale (HADS-D) cut-off values of 5-15 (points represent cut-off values of 5 (right) to 15 (left)) among semi-structured diagnostic interviews, fully structured diagnostic interviews, and the Mini International Neuropsychiatric Interview (MINI), with AUC values

with our previous results in IPDMAs that assessed the accuracy of other screening tools, including the Patient Health Questionnaire-9[38] and the Edinburgh Postnatal Depression Scale.[39] Ideally, data from studies that used different reference standards could be combined but the different diagnostic characteristics of different types of interviews is a barrier.[30-33] Future studies are needed to develop different approaches to combine data even when different reference standards are used while capitalising on the substantial amount of individual participant data we have collected. Also, head-to-head studies for commonly used screening tools for depression are needed.

In this study, we found that a HADS-D cut-off value of seven or higher maximised combined sensitivity and specificity among primary studies that used semi-structured interviews. A cut-off value of eight or higher had similar combined sensitivity and specificity but was less sensitive and more specific. Other cut-off values could be used in clinical practice or trials to prioritise sensitivity or specificity. For example, if a clinician intends to use the HADS-D only to identify medically ill patients with high depressive symptom levels, higher cut-off values could be used to reduce false positives. On the other hand, if the HADS-D is used to capture all patients who might meet the diagnostic criteria based on further assessment, lower cut-off values could be used to avoid false negatives. Based on the results of our IPDMA, a web based knowledge translation tool (depressionscreening100.com/hads-d) was developed to estimate expected numbers of positive screens, and true and false screening outcomes. Clinicians and researchers who consider screening for depression with the HADS-D can refer to this tool.

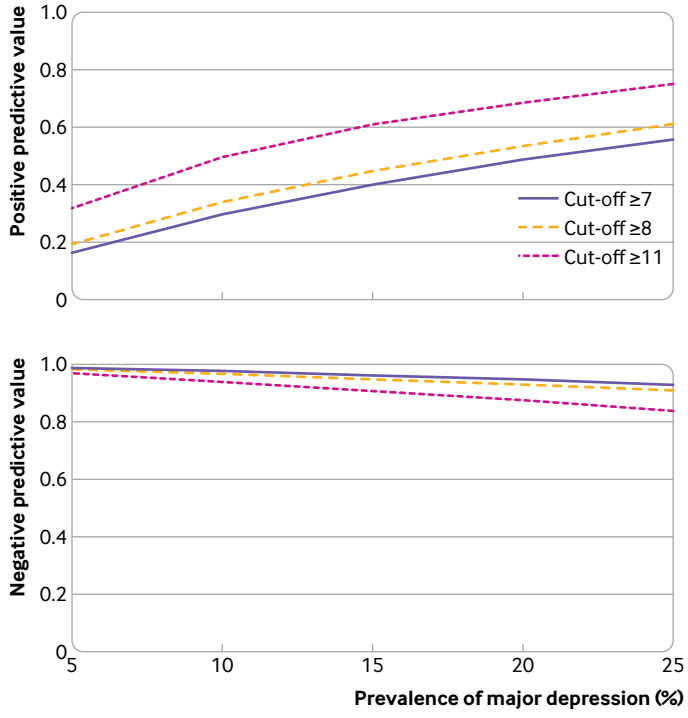

Fig 3 | Nomograms of positive and negative predictive values for cut-off values of seven or higher, eight or higher, and 11 or higher of the Hospital Anxiety and Depression Scale Depression subscale (HADS-D) for a prevalence of major depression of 5-25%, with semi-structured diagnostic interviews as reference standard

Recommendations for routine screening for depression in primary care differ by country. Screening is not directly recommended in the UK, but recommendations from NICE suggest that clinicians might consider asking screening questions.[7 22] Screening is recommended in the United States[13] but not in Canada. The Canadian Task Force on Preventive Health Care has raised concerns about the lack of evidence from trials showing benefit, and about adverse outcomes and the use of scarce healthcare resources.[65] In some countries, specific recommendations for screening in people with a physical illness have been made.[66] Well designed trials evaluating the effects of screening across a range of cut-off scores are needed to determine if screening improves mental health outcomes while minimising harm and unnecessary use of resources. Ideally, trials would also help us to understand how different cut-off values on the HADS-D might influence results.

Strengths and limitations of the study

To our knowledge, ours is the first IPDMA that has analysed the diagnostic accuracy of the HADS-D to detect major depression. Strengths of the study include the large sample size, inclusion of results from all cut-off values from all studies (rather than only those published), and assessment of the accuracy of the HADS-D separately across reference standards and by participant subgroups.

This study has several limitations. Firstly, because of the time required in IPDMAs for updating searches, obtaining primary datasets, and cleaning and synthesising new datasets, the search was not updated after 25 October 2018, and so more recently published studies were not included. Secondly, primary data from 66 of 158 published eligible datasets (42%) were not included but 50 (76%) of these studies did not publish eligible estimates of diagnostic accuracy (supplementary table B2).

Thirdly, moderate heterogeneity was found across studies, which improved in most cases when subgroups were considered. Methods for estimating and interpreting heterogeneity in meta-analyses of test accuracy are not well established, and no recognised guidelines exist for interpretating the results of the quantitative metrics that we used. High heterogeneity in meta-analyses of test accuracy studies is common.[36 60] Subgroup analyses could not be conducted based on medical comorbidities, with the exception of a diagnosis of cancer, as specified in the study protocol, because other than the subgroup of cancer patients (5805 participants with a diagnosis of cancer, 27 studies involved cancer patients), none of the disease based subgroups were large enough. Subgroup analyses on country and language also could not be conducted because many countries and languages were represented in a few primary studies. For example, studies that used 19 different languages were included in our IPDMA but most were represented in only a small number of studies.

Fourthly, many studies included in the meta-analysis did not explicitly exclude participants who might have already been diagnosed or receiving care for depression, although we found no statistically significant differences between analyses of participants verified as not currently diagnosed or receiving treatment for depression and analyses of all participants, including those without this information. Fifthly, studies in the IPDMA were categorised based on the interview conducted but interviews might not have been consistently carried out in the way intended. Among 57 studies that used semi-structured interviews, 11 were rated as unclear for the qualification of the person who conducted the interview. The use of non-qualified interviewers might have reduced differences in estimates of accuracy across reference standard categories. Nonetheless, accuracy was highest when compared with semi-structured interview at a cut-off value of seven or higher, although the difference from other reference standards was not statistically significant across all cut-off values. Lastly, sensitivity analyses including only studies with a low risk of bias rating across all QUADAS-2 domains could not be conducted because of the small number of studies with all low ratings.

Conclusions

In this IPDMA, we found that combined sensitivity and specificity for the HADS-D was maximised at a cut-off value of seven or higher, which was similar to the summed values for a cut-off value of eight or higher. Accuracy was not significantly different across all cut-off values based on reference standards or participant characteristics, including age, sex, diagnosis of cancer, human development index levels, and recruitment setting of participants. Clinicians and researchers who consider screening for depression with the HADS-D can refer to depressionscreening100.com/hads-d to identify alternative cut-off values if sensitivity or specificity is a priority in clinical practice or trials. Well designed trials are needed to determine whether screening with the HADS-D improves mental health outcomes and minimises harm and use of resources.

AUTHOR AFFILIATIONS

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, QC, Canada

[2]Department of Psychiatry, McGill University, Montréal, QC, Canada

[3]Centre for Prognosis Research, School of Primary, Community and Social Care Medicine, Keele University, Staffordshire, UK

[4]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, Canada

[5]Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, QC, Canada

[6]Department of Medicine, McGill University, Montréal, QC, Canada

[7]Department of Psychology, McGill University, Montréal, QC, Canada

[8]Department of Educational and Counselling Psychology, McGill University, Montréal, QC, Canada

[9]Biomedical Ethics Unit, McGill University, Montréal, QC, Canada

We thank Carlos E da Rocha e Silva and Anna P B M Braeken for contributing primary datasets.

Contributors: YW, BLevis, ABenedetti, and BDT were responsible for the study conception and design. YW, BL, YS, CH, AK, DN, PMB, ZN, and BDT contributed to data extraction, coding, evaluation of included studies, and data synthesis. YW, BLevis, ABenedetti, and BDT contributed to data analysis and interpretation. YW, AB, and BDT drafted the manuscript. ABenedetti and BDT contributed equally as co-senior authors and are the guarantors; they had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analyses. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. Members of the DEPRESSD HADS Group contributed: to data extraction, coding, and synthesis: KER, DBR, MA, XWY, MI, MJC, and NS; to design and conduct of database searches: JTB and LAK; as members of the DEPRESSD Steering Committee, including conception and oversight of collaboration: PC, SG, JPAI, SBP, SM, and RCZ; as a knowledge user consultant: MHenry, ZI, CGL, NDM, MTonelli; by contributing included datasets: SA-A, KRB, ABeraldi, CNB, BB, NB-D, ABunevicius, CC, GCarter, C-KC, GCheung, KC, GC-R, DC, ED, FMD, JDS, MD, MGD, AF, PPF, FHF, AJF, MF, PG, MG, LG, MHärter, AH, JJ, NJ, MJ, MKeller, S-WK, MKjærgaard, SK, H-HK, LKRK, YL, ML, WLL, AWL, BL, UFM, RAM, RM-S, LMassardo, YM, AM, IM, LMisery, RN, CJN, CGN, MLO, SJO, AÖ, AP, JAP, JP, LP, JLP, FP, TJQ, SER, KR, NR, SGR-H, AGR, RS-G, RMS, MPJS, MScherer, MLS, VSC, JShaaban, LSharpe, MSharpe, SSimard, SSinger, LStafford, JStone, NAS, SSultan, ALT, IT, MTschorn, K-YT, AT, MWagner, JWalker, MWalterfang, L-JW, SBW, JWhite, LJW, and L-YW. All authors, including group authors, provided a critical review and approved the final manuscript.

Funding: The study was funded by the Canadian Institutes of Health Research (CIHR, KRS-140045 and PCG-155468). YW and BL were supported by Fonds de recherche du Québec-Santé (FRQ-S) Postdoctoral Training Fellowships. DN was supported by GR Caverhill Fellowship from the Faculty of Medicine, McGill University. PMB was supported by a studentship from the Research Institute of the McGill University Health Centre. AB was supported by a FRQ-S researcher salary award. BDT was supported by a Tier 1 Canada Research Chair. DBR was supported by a Vanier Canada Graduate Scholarship. The primary studies by Scott et al, Amoozegar et al, and Prisnie et al were supported by the University of Calgary Cumming School of Medicine, Alberta Health Services, and the Hotchkiss Brain Institute. SBP was supported by a Senior Health Scholar Award from Alberta Innovates, Health Solutions. The primary study by Marrie et al was supported by the CIHR (THC-135234), Crohn's and Colitis Canada, a Research Manitoba Chair, and the Waugh Family Chair in Multiple Sclerosis (to RAM). The primary study by Bernstein et al was supported by the CIHR (THC-135234) and Crohn's and Colitis Canada. CNB was supported in part by the Bingham Chair in Gastroenterology. RAM was supported by the Waugh Family Chair in Multiple Sclerosis and the Research Manitoba Chair. The primary study by Butnoriene et al was supported by a grant from the Research Council of Lithuania (LIG-03/2011). Jurate Butnoriene, PhD, who did the data collection and analysis as part of her PhD thesis for the primary study by Butnoriene et al, passed away and could not participate in this project. Dr Robertas Bunevicius, MD, PhD (1958-2016) was the principal investigator of the primary studies by Butnoriene et al and Bunevicius et al, but passed away and could not participate in this project. The primary study by Chen et al was supported by the National Science Council, Taiwan (NSC 96-2314-B-182A-090-MY2). The primary study by Cheung et al was supported by the Waikato Clinical School, University of Auckland, the Waikato Medical Research Foundation and the Waikato Respiratory Research Fund. The primary study by Costa-Requena et al was supported by the Catalan Agency for Health Technology Assessment and Research (No 102/19/2004). The primary study by Cukor et al was supported in part by a Promoting Psychological Research and Training on Health-Disparities Issues at Ethnic MinorityServing Institutions Grants (ProDIGs) awarded to DC from the American Psychological Association. The primary study by De la Torre et al was supported by a Research Grant "Ramón Carrillo-Arturo Oñativa for Multicentric Studies" (2015) from the commission "Salud Investiga" of the Ministry of Health and Social Action of Argentina (grant No 1853). The primary study by De Souza et al was supported by Birmingham and Solihull Mental Health Foundation Trust. The primary study by Dorow et al was supported by the German Federal Ministry of Education and Research (grant/award No 01GY1155A). The primary study by Douven et al was supported by Maastricht University, Health Foundation Limburg and the Adriana van Rinsum-Ponsen Stichting. The primary study by Honarmand et al was supported by a grant from the Multiple Sclerosis Society of Canada. The primary study by Fischer et al was supported as part of the RECODEHF study by the German Federal Ministry of Education and Research (01GY1150). The primary study by Gagnon et al was supported by the Drummond Foundation and the Department of Psychiatry, University Health Network. The primary study by Akechi et

al was supported in part by a Grant-in-Aid for Cancer Research (11-2) from the Japanese Ministry of Health, Labour, and Welfare and a Grant-in-Aid for Young Scientists (B) from the Japanese Ministry of Education, Culture, Sports, Science, and Technology. The primary study by Kugaya et al was supported in part by a Grant-in-Aid for Cancer Research (9-31) and the Second-Term Comprehensive 10-year Strategy for Cancer Control from the Japanese Ministry of Health, Labour, and Welfare. The primary study Ryan et al was supported by the Irish Cancer Society (grant CRP08GAL). The primary study by Grassi et al was supported by the European Commission DG Health and Consumer Protection (agreement with the University of Ferrara-SI2.307317 2000CVGG2-026), the University of Ferrara, and the Fondazione Cassa di Risparmio di Ferrara. The primary study by Härter et al was supported by the Federal Ministry of Education and Research, the Federation of German Pension Insurance Institutes, and the Freiburg/Bad Saeckingen Rehabilitation Research Network (grant 01 GD 9802/4). The primary study by Keller et al was supported by the Medical Faculty of the University of Heidelberg (grant No 175/2000). The primary study by Kang et al was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (2009-0087344), and was supported by a grant of the Korea Health 21 R&D, Ministry of Health and Welfare, Republic of Korea (A102065). The primary study by Jang et al was supported by a grant from the Korea Health 21 R&D, Ministry of Health and Welfare, Republic of Korea. The primary study by Love et al (2004) was supported by the Kathleen Cuningham Foundation (National Breast Cancer Foundation), the Cancer Council of Victoria and the National Health and Medical Research Council. The primary study by Love et al (2002) was supported by a grant from the Bethlehem Griffiths Research Foundation. The primary study by Löwe et al was supported by the medical faculty of the University of Heidelberg, Germany (project 121/2000). The primary study by Navines et al was supported in part by grants from the Instituto de Salud Carlos III (EO PI08/90869) and (PSIGEN-VHC Study: FIS-E08/00268). The primary study by Massardo et al was supported by Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) grant No PFB12/2007 and Fondo Nacional de Desarrolo Científico y Tecnológico (FONDECYT; grant No 1110849). The primary study by Matsuoka et al was supported by the Japanese Ministry of Health, Labour, and Welfare through Research on Psychiatric and Neurological Disease and Mental Health (16190501, 19230701, and 20300701). The primary study by Hartung et al was supported by the German Cancer Aid within the psychosocial oncology funding priority program (grant No 107465). The primary study by Consoli et al was supported by grants from the French Society of Dermatology and the University Hospital of Saint Etienne. The primary study by McFarlane et al was supported by an Australian Government National Health and Medical Research Council program grant. MLO was supported by grants from NHMRC Program (1073041) during the conduct of the study. The primary study by O'Rourke et al was supported by the Scottish Home and Health Department, Stroke Association, and Medical Research Council. The primary study by Sanchez-Gistau et al was supported by a grant from the Ministry of Health of Spain (PI040418) and in part by the Catalonia Government, DURSI 2009SGR1119. The primary study by Gould et al was supported by the Transport Accident Commission Grant. The primary study by Bayon-Perez et al was supported by a grant from the Instituto de Investigación Hospital 12 de Octubre (i+12). FP was an investigator from the Intensification of Research Activity Program of the Instituto de Investigación Hospital 12 de Octubre (i+12) during the conduct of the study. The primary study by Lees et al was supported by a "start-up" research grant from the British Geriatric Society, Scotland. The primary study by Reme et al was supported by the Research Council of Norway. The primary study by Rooney et al was supported by the NHS Lothian Neuro-Oncology Endowment Fund. The primary study by Schwarzbold et al was supported by PRONEX Program (NENASC Project) and PPSUS Program of Fundaçao de Amparo a esquisa e Inovacao do Estado de Santa Catarina (FAPESC) and the National Science and Technology Institute for Translational Medicine (INCT-TM). The primary study by Azah et al was supported by Universiti Sains Malaysia. The primary studies by Patel et al (2010 and 2011) were supported by the University of Sydney Cancer Research Fund. The primary study by Simard et al was supported by IDEA grants from the Canadian Prostate Cancer Research Initiative and the Canadian Breast Cancer Research Alliance, as well as a studentship from the CIHR. The primary study by Singer et al (2009) was supported by a grant from the German Federal Ministry for Education and Research (No 01ZZ0106). The primary study by Singer et al (2008) was supported by grants from the German Federal Ministry for Education and Research (No 7DZAIQTX) and of the University of Leipzig (No formel. 1-57). The primary study by Meyer et al was supported by the Federal Ministry of Education and Research (BMBF). The primary study by Stafford et al (2014) was supported in part by seed funding from the Western and Central Melbourne Integrated Cancer Service. The primary study by Stafford et al (2007) was supported by the University of Melbourne. The primary study by Stone et al was supported by the Medical Research Council, UK and Chest Heart and Stroke, Scotland. The primary study by Phan et al was supported by the Government of Western Australia, Department of Health (grant No G1000794). The primary study by de Oliveira et al was supported by CNPq and Fapemig, Brazil. The primary study by Pedroso et al (2018) was supported by FAPEMIG (APq-03539-13). The primary study by Pedroso et al (2016) was supported by Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig) (APq-03539-13). The primary study by Tiringer et al was supported by the Hungarian Research Council (ETT 395). The primary study by Tschorn et al is part of the study "CDCare-Supply of patients with coronary artery disease and comorbid depression: a patient oriented needs analysis." CDCare was funded by the Federal Ministry of Education and Research (BMBF; 01GY1154). The primary study by Turner et al was supported by a bequest from Jennie Thomas through Hunter Medical Research Institute. The primary study by Walterfang et al was supported by Melbourne Health. The primary study by Lee et al (2017) was supported by a grant from the Kaohsiung Chang Gung Memorial Hospital, Taiwan (CMRPG8A0581). The primary study by Lee et al (2016) was supported by a grant from Kaohsiung Chang Gung Memorial Hospital, Taiwan (CMRPG891321). The primary study by Sia et al (PIs: Pasco and Williams) was supported by the Victorian Health Promotion Foundation (ID 91-0095) and the National Health and Medical Research Council (ID 628582; 299831; 251638; 509103; 1026265; 009367; 1104438). No other authors reported funding for primary studies or for their work on this study. No funder had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Competing interests: All authors have completed the ICJME uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years with the following exceptions: ZI declares that he has received personal fees from Avanir, Janssen, Lundbeck, Otsuka, and Sunovion, outside the submitted work. MTonelli declares that he has received a grant from Merck Canada, outside the submitted work. CNB declares that he has consulted for Abbvie Canada, Amgen Canada, Bristol Myers Squibb Canada, Roche Canada, Janssen Canada, Pfizer Canada, Sandoz Canada, Takeda Canada, and Mylan Pharmaceuticals. He has also received unrestricted educational grants from Abbvie Canada, Janssen Canada, Pfizer Canada, and Takeda Canada; as well as been on speaker's bureau of Abbvie Canada, Janssen Canada, Takeda Canada, and Medtronic Canada, all outside the submitted work. AF reports that he received speaker's honorariums from Biogen, Sanofi-Genzyme, Merck-Serono, Novartis, and Roche, and is on the advisory board for Akili Interactive, outside the submitted work; he has also received royalties from the Cambridge University Press for the Clinical Neuropsychiatry of Multiple Sclerosis, 2nd edition. BLöwe declares that the primary study by Löwe et al was supported by unrestricted educational grants from Pfizer, Germany. RAM declares that she has conducted clinical trials for Sanofi Aventis, outside the submitted work. YM declares that he has received personal fees from Mochida, Pfizer, Eli Lilly, Morinaga Milk, and NTT Data, outside the submitted work. SSinger declares that she has received personal fees from Lilly, BMS, and Pfizer, outside the submitted work. JStone declares that he has received personal fees from UptoDate, outside the submitted work. SSultan declares funding from Sanofi-Aventis Corporation, during conduct of the primary study. All authors declare no other relationships or activities that could appear to have influenced the submitted work. No funder had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Ethical approval: As the study involved secondary analysis of anonymised previously collected data, the research ethics committee of the Jewish General Hospital declared that this project did not require research ethics approval. For each included dataset, however, the authors confirmed that the original study received ethics approval and that all patients provided informed consent.

Data sharing: Requests to access data should be made to the corresponding author.

The manuscript's guarantor affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Dissemination to participants and related patient and public communities: There are no plans to disseminate the results of the research to study participants or the relevant patient community. A web based knowledge translation tool, intended for clinicians (the end-users of the HADS screening tool), however, is available at depressionscreening100.com/hads-d. The tool allows clinicians to estimate the expected number of positive screens and true and false screening outcomes based on study results.

Provenance and peer review: Not commissioned; externally peer-reviewed.

1 Rudisch B, Nemeroff CB. Epidemiology of comorbid coronary artery disease and depression. *Biol Psychiatry* 2003;54:227-40. doi:10.1016/S0006-3223(03)00587-0
2 Thombs BD, Bass EB, Ford DE, et al. Prevalence of depression in survivors of acute myocardial infarction. *J Gen Intern Med* 2006;21:30-8. doi:10.1111/j.1525-1497.2005.00269.x
3 Evans DL, Charney DS, Lewis L, et al. Mood disorders in the medically ill: scientific review and recommendations. *Biol Psychiatry* 2005;58:175-89. doi:10.1016/j.biopsych.2005.05.001
4 Dickens C, Creed F. The burden of depression in patients with rheumatoid arthritis. *Rheumatology (Oxford)* 2001;40:1327-30. doi:10.1093/rheumatology/40.12.1327
5 Ali S, Stone MA, Peters JL, Davies MJ, Khunti K. The prevalence of co-morbid depression in adults with Type 2 diabetes: a systematic review and meta-analysis. *Diabet Med* 2006;23:1165-73. doi:10.1111/j.1464-5491.2006.01943.x
6 Fann JR, Thomas-Rich AM, Katon WJ, et al. Major depression after breast cancer: a review of epidemiology and treatment. *Gen Hosp Psychiatry* 2008;30:112-26. doi:10.1016/j.genhosppsych.2007.10.008
7 National Institute for Clinical Excellence. *Depression in adults with a chronic physical health problem: recognition and management. Clinical guideline.* [CG91] National Institute for Clinical Excellence, 2010.
8 Duhoux A, Fournier L, Gauvin L, Roberge P. What is the association between quality of treatment for depression and patient outcomes? A cohort study of adults consulting in primary care. *J Affect Disord* 2013;151:265-74. doi:10.1016/j.jad.2013.05.097
9 Duhoux A, Fournier L, Nguyen CT, Roberge P, Beveridge R. Guideline concordance of treatment for depressive disorders in Canada. *Soc Psychiatry Psychiatr Epidemiol* 2009;44:385-92. doi:10.1007/s00127-008-0444-8
10 Mojtabai R. Clinician-identified depression in community settings: concordance with structured-interview diagnoses. *Psychother Psychosom* 2013;82:161-9. doi:10.1159/000345968
11 Mojtabai R, Olfson M. Proportion of antidepressants prescribed without a psychiatric diagnosis is growing. *Health Aff (Millwood)* 2011;30:1434-42. doi:10.1377/hlthaff.2010.1024
12 Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet* 2009;374:609-19. doi:10.1016/S0140-6736(09)60879-5
13 Siu AL, Bibbins-Domingo K, Grossman DC, et al, US Preventive Services Task Force (USPSTF). Screening for depression in adults: US Preventive Services Task Force recommendation statement. *JAMA* 2016;315:380-7. doi:10.1001/jama.2015.18392
14 Palmer SC, Coyne JC. Screening for depression in medical care: pitfalls, alternatives, and revised priorities. *J Psychosom Res* 2003;54:279-87. doi:10.1016/S0022-3999(02)00640-2
15 Gilbody S, Sheldon T, Wessely S. Should we screen for depression?*BMJ* 2006;332:1027-30. doi:10.1136/bmj.332.7548.1027
16 Thombs BD, Coyne JC, Cuijpers P, et al. Rethinking recommendations for screening for depression in primary care. *CMAJ* 2012;184:413-8. doi:10.1503/cmaj.111035
17 Thombs BD, Ziegelstein RC. Does depression screening improve depression outcomes in primary care?*BMJ* 2014;348:g1253. doi:10.1136/bmj.g1253
18 Thombs BD, Ziegelstein RC, Roseman M, Kloda LA, Ioannidis JP. There are no randomized controlled trials that support the United States Preventive Services Task Force Guideline on screening for depression in primary care: a systematic review. *BMC Med* 2014;12:13. doi:10.1186/1741-7015-12-13
19 National Institute for Clinical Excellence. *Improving supportive and palliative care for adults with cancer. Cancer service guideline.* [CSG4] National Institute for Clinical Excellence, 2004.
20 Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;67:361-70. doi:10.1111/j.1600-0447.1983.tb09716.x
21 Meader N, Mitchell AJ, Chew-Graham C, 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. doi:10.3399/bjgp11X613151
22 National Institute for Health and Care Excellence. *Depression in adults. Quality standard.* [QS8] National Institute for Health and Care Excellence, 2011.
23 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. doi:10.1016/j.jpsychores.2010.04.006
24 Vodermaier A, Millman RD. Accuracy of the Hospital Anxiety and Depression Scale as a screening tool in cancer patients: a systematic review and meta-analysis. *Support Care Cancer* 2011;19:1899-908. doi:10.1007/s00520-011-1251-4
25 Mitchell AJ, Meader N, Symonds P. Diagnostic validity of the Hospital Anxiety and Depression Scale (HADS) in cancer and palliative settings: a meta-analysis. *J Affect Disord* 2010;126:335-48. doi:10.1016/j.jad.2010.01.067
26 Thombs BD, Rice DB. Sample sizes and precision of estimates of sensitivity and specificity from primary studies on the diagnostic accuracy of depression screening tools: a survey of recently published studies. *Int J Methods Psychiatr Res* 2016;25:145-52. doi:10.1002/mpr.1504
27 Levis B, Benedetti A, Levis AW, et al. Selective Cutoff Reporting in Studies of Diagnostic Test Accuracy: A Comparison of Conventional and Individual-Patient-Data Meta-Analyses of the Patient Health Questionnaire-9 Depression Screening Tool. *Am J Epidemiol* 2017;185:954-64. doi:10.1093/aje/kww191
28 Thombs BD, Arthurs E, El-Baalbaki G, Meijer A, Ziegelstein RC, Steele RJ. Risk of bias from inclusion of patients who already have diagnosis of or are undergoing treatment for depression in diagnostic accuracy studies of screening tools for depression: systematic review. *BMJ* 2011;343:d4825. doi:10.1136/bmj.d4825
29 Rice DB, Thombs BD. Risk of bias from inclusion of currently diagnosed or treated patients in studies of depression screening tool accuracy: A cross-sectional analysis of recently published primary studies and meta-analyses. *PLoS One* 2016;11:e0150067. doi:10.1371/journal.pone.0150067
30 Levis B, Benedetti A, Riehm KE, et al. Probability of major depression diagnostic classification using semi-structured versus fully structured diagnostic interviews. *Br J Psychiatry* 2018;212:377-85. doi:10.1192/bjp.2018.54
31 Levis B, McMillan D, Sun Y, et al. Comparison of major depression diagnostic classification probability using the SCID, CIDI, and MINI diagnostic interviews among women in pregnancy or postpartum: An individual participant data meta-analysis. *Int J Methods Psychiatr Res* 2019;28:e1803. doi:10.1002/mpr.1803
32 Wu Y, Levis B, Sun Y, et al. Probability of major depression diagnostic classification based on the SCID, CIDI and MINI diagnostic interviews controlling for Hospital Anxiety and Depression Scale - Depression subscale scores: An individual participant data meta-analysis of 73 primary studies. *J Psychosom Res* 2020;129:109892. doi:10.1016/j.jpsychores.2019.109892
33 Wu Y, Levis B, Ioannidis JPA, Benedetti A, Thombs BDDEPRESSion Screening Data (DEPRESSD) Collaboration. Probability of Major Depression Classification Based on the SCID, CIDI, and MINI Diagnostic Interviews: A Synthesis of Three Individual Participant Data Meta-Analyses. *Psychother Psychosom* 2021;90:28-40. doi:10.1159/000509283
34 Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;340:c221. doi:10.1136/bmj.c221
35 Thombs BD, Benedetti A, Kloda LA, et al. Diagnostic accuracy of the Depression subscale of the Hospital Anxiety and Depression Scale (HADS-D) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *BMJ Open* 2016;6:e011913. doi:10.1136/bmjopen-2016-011913
36 Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. *BMJ* 2020;370:m2632. doi:10.1136/bmj.m2632
37 Stewart LA, Clarke M, Rovers M, et al, PRISMA-IPD Development Group. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. *JAMA* 2015;313:1657-65. doi:10.1001/jama.2015.3656

38 Levis B, Benedetti A, Thombs BDDEPRESsion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ* 2019;365:l1476. doi:10.1136/bmj.l1476
39 Levis B, Negeri Z, Sun Y, Benedetti A, Thombs BDDEPRESsion Screening Data (DEPRESSD) EPDS Group. Accuracy of the Edinburgh Postnatal Depression Scale (EPDS) for screening to detect major depression among pregnant and postpartum women: systematic review and meta-analysis of individual participant data. *BMJ* 2020;371:m4022. doi:10.1136/bmj.m4022
40 *Diagnostic and statistical manual of mental disorders: DSM-III.* 3rd ed, revised. American Psychiatric Association, 1987.
41 *Diagnostic and statistical manual of mental disorders: DSM-IV.* 4th ed. American Psychiatric Association, 1994.
42 *Diagnostic and statistical manual of mental disorders: DSM-IV.* 4th ed, text revised. American Psychiatric Association, 2000.
43 *Diagnostic and statistical manual of mental disorders: DSM-V.* 5th ed. American Psychiatric Association, 2013.
44 WHO. *The ICD-10 Classifications of Mental and Behavioural Disorders—Clinical Descriptions and Diagnostic Guidelines.* World Health Organization, 1992.
45 PRESS – Peer Review of Electronic Search Strategies. 2015 Guideline Explanation and Elaboration (PRESS E&E). Ottawa: CADTH 2016.
46 United Nations Development Programme. Human Development Reports. http://hdr.undp.org/en/content/human-development-index-hdi.
47 Whiting PF, Rutjes AW, Westwood ME, et al, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529-36. doi:10.7326/0003-4819-155-8-201110180-00009
48 First MB. *Structured Clinical Interview for the DSM (SCID).* John Wiley & Sons, Inc, 1995.
49 World Health Organization. *Schedules for clinical assessment in neuropsychiatry: manual.* Amer Psychiatric Pub Inc, 1994.
50 Endicott J, Spitzer RL. Schedule for affective disorders and schizophrenia (SADS). *Acta Psychiatr Belg* 1987;87:361-516.
51 Clarke DM, Smith GC, Herrman HE, McKenzie DP. Monash Interview for Liaison Psychiatry (MILP). Development, reliability, and procedural validity. *Psychosomatics* 1998;39:318-28. doi:10.1016/S0033-3182(98)71320-9
52 Robins LN, Wing J, Wittchen HU, et al. The Composite International Diagnostic Interview. An epidemiologic Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Arch Gen Psychiatry* 1988;45:1069-77. doi:10.1001/archpsyc.1988.01800360017003
53 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. doi:10.1001/archpsyc.1981.01780290015001
54 Lecrubier Y, Sheehan DV, Weiller E, et al. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *Eur Psychiatry* 1997;12:224-31. doi:10.1016/S0924-9338(97)83296-8
55 Sheehan DV, Lecrubier Y, Sheehan KH, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *Eur Psychiatry* 1997;12:232-41. doi:10.1016/S0924-9338(97)83297-X
56 Brugha TS, Jenkins R, Taub N, Meltzer H, Bebbington PE. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychol Med* 2001;31:1001-13. doi:10.1017/S0033291701004184
57 Brugha TS, Bebbington PE, Jenkins R. A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. *Psychol Med* 1999;29:1013-20. doi:10.1017/S0033291799008880
58 Nosen E, Woody SR. Diagnostic Assessment in Research. In: *McKay D. Handbook of research methods in abnormal and clinical psychology.* Sage, 2008: 109-24.
59 Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Stat Med* 2008;27:6111-36. doi:10.1002/sim.3441
60 Macaskill P, Gatsonis C, Deeks JJ, et al. Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.* Cochrane Collaboration, 2010. https://methods.cochrane.org/sdt/.
61 Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539-58. doi:10.1002/sim.1186
62 R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
63 RStudio Team. (2020). RStudio: Integrated Development for R. RStudio, Boston, MA, USA. https://www.rstudio.com/.
64 Bates D, Maechler M, Bolker B, et al. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* 2015;67:1-48. doi:10.18637/jss.v067.i01.
65 Joffres M, Jaramillo A, Dickinson J, et al, Canadian Task Force on Preventive Health Care. Recommendations on screening for depression in adults. *CMAJ* 2013;185:775-82. doi:10.1503/cmaj.130403
66 Härter M, Klesse C, Bermejo I, Schneider F, Berger M. Unipolar depression: diagnostic and therapeutic recommendations from the current S3/National Clinical Practice Guideline. *Dtsch Arztebl Int* 2010;107:700-8.

Web appendix 1: Members of the DEPRESSD HADS Group
Web appendix 2: Supplementary material