Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis

Zelalem F Negeri,[1,2] Brooke Levis,[3] Ying Sun,[1] Chen He,[1] Ankur Krishnan,[1] Yin Wu,[1,4] Parash Mani Bhandari,[1,2] Dipika Neupane,[1,2] Eliana Brehaut,[1] Andrea Benedetti,[2,5,6] Brett D Thombs,[1,2,4,5,7,8,9] on behalf of the Depression Screening Data (DEPRESSD) PHQ 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;374:n2183
http://dx.doi.org/10.1136/bmj.n2183

Accepted: 23 August 2021

ABSTRACT

OBJECTIVE

To update a previous individual participant data meta-analysis and determine the accuracy of the Patient Health Questionnaire-9 (PHQ-9), the most commonly used depression screening tool in general practice, for detecting major depression overall and by study or participant subgroups.

DESIGN

Systematic review and individual participant data meta-analysis.

DATA SOURCES

Medline, Medline In-Process, and Other Non-Indexed Citations via Ovid, PsycINFO, Web of Science searched through 9 May 2018.

REVIEW METHODS

Eligible studies administered the PHQ-9 and classified current major depression status using a validated semistructured diagnostic interview (designed for clinician administration), fully structured interview (designed for lay administration), or the Mini International Neuropsychiatric Interview (MINI; a brief interview designed for lay administration). A bivariate random effects meta-analytic model was used to obtain point and interval estimates of pooled PHQ-9 sensitivity and specificity at cut-off values 5-15, separately, among studies that used semistructured diagnostic interviews (eg, Structured Clinical Interview for Diagnostic and Statistical Manual), fully structured interviews (eg, Composite International Diagnostic Interview), and the MINI. Meta-regression was used to investigate whether PHQ-9 accuracy correlated with reference standard categories and participant characteristics.

RESULTS

Data from 44 503 total participants (27 146 additional from the update) were obtained from 100 of 127 eligible studies (42 additional studies; 79% eligible studies; 86% eligible participants). Among studies with a semistructured interview reference standard, pooled PHQ-9 sensitivity and specificity (95% confidence interval) at the standard cut-off value of ≥10, which maximised combined sensitivity and specificity, were 0.85 (0.79 to 0.89) and 0.85 (0.82 to 0.87), respectively. Specificity was similar across reference standards, but sensitivity in studies with semistructured interviews was 7-24% (median 21%) higher than with fully structured reference standards and 2-14% (median 11%) higher than with the MINI across cut-off values. Across reference standards and cut-off values, specificity was 0-10% (median 3%) higher for men and 0-12 (median 5%) higher for people aged 60 or older.

CONCLUSIONS

Researchers and clinicians could use results to determine outcomes, such as total number of positive screens and false positive screens, at different PHQ-9 cut-off values for different clinical settings using the knowledge translation tool at www.depressionscreening100.com/phq.

STUDY REGISTRATION

PROSPERO CRD42014010673.

WHAT IS ALREADY KNOWN ON THIS TOPIC

The Patient Health Questionnaire-9 (PHQ-9) is the most commonly used depression screening tool in primary and general settings, with cut-off value ≥10 used as a standard to identify major depression

A previous individual participant data meta-analysis on PHQ-9 accuracy for detecting major depression included 58 studies (17 357 participants) through February 2015 and found that the PHQ-9 had higher accuracy in comparison with semistructured reference standards than with other reference standards; older age of participants was significantly, although minimally, associated with higher specificity; combined sensitivity and specificity was maximised at the standard cut-off value of ≥10

WHAT THIS STUDY ADDS

Updated searches through May 2018 showed that overall sensitivity and specificity estimates were robust and consistent with previous estimates, even though the sample included an additional 42 studies (27 146 additional participants)

PHQ-9 specificity was slightly better when estimated among only participants confirmed as not already diagnosed or receiving treatment and who would be screened in practice

Across all possible cut-off values, for semistructured interviews, specificity was 0-12% (median 5%) higher for older participants, which contradicts assumptions that screening tools might be less accurate in elderly people

A knowledge translation tool (www.depressionscreening100.com/phq) based on the findings from this study can be used to generate screening outcomes for different cut-off values based on local assumptions about prevalence

Introduction

Depression accounts for more years of healthy life lost than any other medical condition.[1-4] Screening has been recommended to identify people with unrecognised depression by the United States Preventive Services Task Force[5] but not the Canadian Task Force on Preventive Health Care[6] or the UK National Screening Committee.[7] Depression symptom

questionnaires can be used for many purposes, including assessing and discussing symptoms with patients who might be unsure if they have depression, monitoring for treatment response or relapse detection, and depression screening.[8] Depression screening involves administering a questionnaire with a prespecified cut-off value to classify patients who have not been otherwise identified as possibly depressed as positive or negative screens and further assessing those with positive screens to determine whether major depression criteria are met.[8-13]

The nine item Patient Health Questionnaire-9 (PHQ-9)[14-16] is recommended for screening by the US Preventive Services Task Force and others.[11 17 18] Items, scored 0-3, reflect how often each of the nine Diagnostic and Statistical Manual (DSM) major depression symptoms[19-22] have bothered respondents in the past two weeks. The standard cut-off value for detecting major depression is ≥10.[14-16 23-25]

A previous individual participant data meta-analysis (IPDMA; 58 studies, 17 357 participants)[25] conducted by our collaboration compared PHQ-9 accuracy with semistructured diagnostic interviews, fully structured diagnostic interviews, and the Mini International Neuropsychiatric Interview (MINI), separately, owing to important differences in the characteristics and performance of different types of diagnostic interviews.[26-29] Semistructured interviews are designed for administration by trained clinicians to replicate clinical diagnostic procedures as closely as possible in a research setting, whereas fully structured interviews are designed for lay administration with no clinical judgment required.[30-33] The MINI is a brief, fully structured interview designed for rapid administration and to be overinclusive.[34 35] The previous IPDMA found that compared with semistructured interviews, PHQ-9 ≥10 had sensitivity of 0.88 (95% confidence interval 0.83 to 0.92), and specificity of 0.85 (0.82 to 0.88). A statistically significant, but minimal, difference in sensitivity by age was found, but no other differences by participant or study level subgroups.

Our objective was to update the previous PHQ-9 IPDMA, using an updated and larger dataset, to firstly, assess PHQ-9 screening accuracy compared with semistructured (primary analysis), fully structured (other than MINI), and MINI diagnostic interviews, separately, and, secondly, investigate accuracy by participant and study characteristic subgroups. In this update, we obtained data from 42 additional studies (27 146 participants) since our previous PHQ-9 IPDMA.[25]

Methods

The IPDMA was registered in PROSPERO international prospective register of systematic reviews (CRD42014010673), and a protocol was published.[36] We followed reporting guidance from Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) of diagnostic test accuracy[37] and PRISMA of individual participant data.[38] Methods were consistent with those of the previous IPDMA.[25] Separate IDPMAs have been published for the PHQ-2[39] and PHQ-8.[40]

Study eligibility

Eligible datasets were sought from studies in any language that included diagnostic classification for current major depressive disorder or major depressive episode based on DSM[19-22] or ICD (international classification of diseases)[41] criteria, using a validated semistructured or fully structured interview conducted within two weeks of PHQ-9 administration, among participants aged 18 or older not recruited from youth or psychiatric settings or because they were already identified as having depression symptoms. Patients from psychiatric settings or those already identified as having depression symptoms were excluded because screening is carried out to identify patients with unrecognised depression.

Datasets where not all participants were eligible were included if primary data allowed selection of eligible participants. For defining major depression, if results from both DSM and ICD or both major depressive disorder and major depressive episode were provided, we prioritised DSM over ICD and prioritised major depressive episode over major depressive disorder. This procedure was followed because DSM classification was used in almost all studies and because screening is done to detect depressive episodes; additional interviews are needed to determine if an episode is related to major depressive disorder, bipolar disorder, or persistent depressive disorder.

Database searches and study selection

Initial and updated Medline, Medline In-Process, and Other Non-Indexed Citations via Ovid, PsycINFO, and Web of Science searches, using peer reviewed[42] search strategies (supplementary methods A), covered 1 January 2000, through February 2015 (initial search) and 9 May 2018 (updated search). The initial search began in 2000 because the PHQ-9 was published in 2001.[14] Additionally, reference lists of relevant reviews were searched, and contributing authors were queried about non-published studies. Search results were uploaded into RefWorks (RefWorks-COS, Bethesda, MD, USA), and, after deduplication, unique citations were uploaded into DistillerSR (Evidence Partners, Ottawa, Canada).

Two independent investigators reviewed titles and abstracts. If either investigator deemed a study potentially eligible, a full text review was carried out by two investigators independently with disagreements resolved by consensus, consulting a third investigator when necessary. Translators were consulted for languages other than those in which team members were fluent.

Data contribution, extraction, and synthesis

Deidentified primary data were requested from investigators with eligible datasets. Corresponding authors were sent up to three emails, as necessary. If not successful, we emailed coauthors and attempted

to contact corresponding authors by phone. Individual participant data that were obtained were transferred to a standard format and merged into a single dataset with study level data. Any discrepancies between published results and raw datasets were resolved in consultation with primary study investigators.

Two investigators independently extracted study level data, including country, recruitment setting (non-medical, primary care, inpatient, outpatient specialty), and diagnostic interview from published articles, consulting a third investigator and querying authors, if necessary. Based on the United Nations' human development index,[43] countries were categorised as "very high," "high," or "low-medium" development based on the index for the year of study publication. Participants' age, sex, major depression interview status, current mental health diagnosis or treatment, and PHQ-9 scores were included in participant level data. Recruitment setting was coded by participant, as two primary studies had more than one recruitment setting. If provided, we used weights from primary studies that implemented weighting to reflect sampling procedures. When primary studies were not weighted, but their sampling procedures warranted weighting, we calculated weights using inverse selection probabilities. Primary studies in which all participants with positive screens and a random subset of participants with negative screens were administered a diagnostic interview, for example, necessitated weighting.

Risk of bias assessment

We assessed risk of bias with the QUality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool, applied by two investigators independently, based on reporting in primary publications. Discussion and consensus were used to resolve any discrepancies with a third investigator involved if necessary. Supplementary methods B present QUADAS-2 coding rules used in the study.[44]

Statistical analyses

Four different sets of analyses were performed. In the first set, we fitted bivariate random effects models to estimate PHQ-9 sensitivity and specificity across cut-off values ≥5 to ≥15 with 95% confidence intervals; we did this separately by studies that used semistructured (primary analysis; Structured Clinical Interview for DSM,[45] Schedules for Clinical Assessment in Neuropsychiatry,[46] Depression Interview and Structured Hamilton[47]), fully structured (Composite International Diagnostic Interview,[48] Clinical Interview Schedule-Revised,[49] Diagnostic Interview Schedule[50]), and MINI[34 35] diagnostic interviews.

In the second set of analyses, we fitted bivariate random effects models to estimate sensitivity and specificity across PHQ-9 cut-off values for only participants known not to be currently diagnosed or receiving treatment for a mental health problem compared with results among all participants. We used this method because, although the PHQ-9 can be used for many purposes (eg, screening, monitoring symptoms during treatment, checking for relapse), screening is done to identify people with previously unrecognised major depression. Already diagnosed or treated patients are sometimes included in primary studies from non-mental health settings but would not be screened in practice.[51 52] Not all primary studies, however, provided data on which participants had previously identified depression, so we compared results with all participants versus results among participants we could confirm were not already diagnosed or receiving treatment. We used a clustered bootstrap approach to construct 95% confidence intervals for differences in sensitivity and specificity at cut-off values of 5-15 between participants not currently diagnosed or receiving treatment for a mental health problem versus all participants[53 54]; 1000 iterations of resampled data were used at the study and subject levels.

In the third set of analyses, we investigated differences in sensitivity and specificity between reference standard categories and participant subgroups using meta-regression. We fitted a one stage multiple meta-regression that interacted reference standard category (reference: semistructured) with PHQ-9 accuracy coefficients (logit(sensitivity) and logit(1−specificity)) and compared results with those obtained from the bivariate random effects model. We fitted additional multiple meta-regression models within each reference standard category by interacting PHQ-9 logit(sensitivity) and logit(1-specificity) with participant characteristics, including continuously measured age, sex (reference: women), country human development index (reference: very high), and participant recruitment setting (reference: primary care). We did not include medical comorbidity in the subgroup analysis because 18 316 of 44 503 (41%) participants did not have comorbidity data, and too few studies had data from single conditions (eg, cancer) to assess by condition. Similarly, there were insufficient studies across categories to analyse results by language or country.

In our fourth set of analyses, following recommendations by Kent et al,[55] we fitted bivariate random effects models among subgroups based on each participant characteristic significantly associated with sensitivity or specificity for all three reference standard categories across all or most cut-off values of 5-15 in the meta-regression models from the third set of analyses. For these analyses, age was dichotomised as less than 60 and 60 or older, as done previously.[25] Primary studies with no patients with major depression or none without depression were excluded in each subgroup analysis because the inclusion of such studies did not allow application of a bivariate random effects model.

For all analyses, bivariate random effects models that considered the correlation between test sensitivity and specificity were fitted to the PHQ-9 data using the Gauss-Hermite adaptive quadrature algorithm[56] and one quadrature point, which is equivalent to the Laplace approximation, to obtain overall sensitivity,

specificity, and associated 95% confidence intervals. We constructed empirical receiver operating characteristic curves for each reference standard based on pooled sensitivity and specificity estimates and calculated areas under the curve.

To quantify statistical heterogeneity across studies in each reference standard category and then separately across participant subgroups within each category, we generated forest plots of sensitivities and specificities. Although there are no well established methods to quantify levels of heterogeneity in meta-analyses of diagnostic test accuracy,[37 57] we quantified heterogeneity by reporting τ^2 (the estimated variances of the random effects for sensitivity and specificity), R (the ratio of the estimated standard deviation of the overall sensitivity (or specificity) from the random effects model to that from the corresponding fixed effects model),[58] and the 95% prediction intervals for the unknown sensitivity and specificity of a new study.

We used sensitivity and specificity estimates at the standard cut-off value ≥10 from our first set of analyses and hypothetical major depression prevalence values of 5-25% to generate nomograms to estimate positive and negative predictive values of the PHQ-9.

As a sensitivity analysis, we fitted multiple meta-regression models based on QUADAS-2 signalling questions for each reference standard category to compare accuracy of results between subgroups based on risk of bias. For these analyses, QUADAS-2 signalling questions were interacted with logit(sensitivity) and logit(1−specificity) for all these questions with a minimum of 100 participants with major depression and 100 without depression among studies categorised as "low" risk of bias versus "high" or "unclear" risk of bias. As an additional sensitivity analysis, we also assessed the effects on the main IPDMA results of including data from eligible studies that did not contribute data but published eligible accuracy data.

We employed R[59] version 4.0.0 and RStudio[60] version 1.2.5042 to run all analyses using the *glmer* function within the *lme4*[61] R package.

Patient and public involvement

No patients were involved in developing the research question, outcome measures, or study design. Since study inception, Dr Sarah Markham joined the DEPRESSD Group as a patient collaborator. She reviewed the draft manuscript.

Results

Search results and dataset inclusion

A total of 9670 unique titles and abstracts were identified from database searches, including the initial and updated searches; 9199 were excluded at the title and abstract review stage and 297 after full text review (supplementary table A), which left 174 articles meeting eligibility criteria. Of these, 56 had duplicate samples. Of the remaining 118 studies, 91 (77%) provided participant data. Nine additional unpublished studies were contributed by authors, which resulted in a total of 100 included studies that provided participant data (number of participants 44 503; number with major depression 4541 (prevalence 10%; fig 1). Of these, 42 studies with 27 146 participants were from the updated search. Supplementary table B shows characteristics of included primary studies and eligible studies that did not provide data.

Table 1 shows participant data by reference standard. Of the 100 included studies, 47 (11 234 participants; 1528 with major depression) used a semistructured interview, 20 (17 167 participants; 1352 major depression) a fully structured interview (other than the MINI), and 33 (16 102 participants; 1661 major depression) the MINI. The Structured Clinical Interview for DSM was the most used semistructured interview (44 of 47 studies), and the Composite International Diagnostic Interview the most used fully structured interview (17 of 20 studies). Table 2 displays participant data by subgrouping variables.

PHQ-9 sensitivity and specificity by reference standard category

Estimates of PHQ-9 sensitivity and specificity for each reference standard category are given in table 3. Cut-off scores of ≥10, ≥8, and ≥8 maximised combined sensitivity and specificity for semistructured, fully structured (excluding MINI), and MINI reference standards, respectively. At the standard cut-off value of ≥10, sensitivity estimates (95% confidence interval) were 0.85 (0.79 to 0.89) for semistructured, 0.64 (0.53 to 0.74) for fully structured, and 0.74 (0.67 to 0.79) for MINI reference standards; corresponding specificity estimates (95% confidence interval) were 0.85 (0.82 to 0.87), 0.88 (0.83 to 0.92). and 0.89 (0.86 to 0.91), respectively. PHQ-9 sensitivity, across all cut-off values, when compared with a semistructured reference standard was 7-24% (median 21%) higher than with fully structured reference standards, and 2-14% (median 11%) higher than with the MINI. PHQ-9 specificity was similar across cut-off values and reference standards. Figure 2 shows receiver operating characteristic plots and area under the curve for each reference standard. Area under the curve was highest when the PHQ-9 was compared with semistructured interviews (0.90), followed by the MINI (0.88) and other fully structured (0.84) reference standards.

Heterogeneity between studies varied from moderate to large, although it was diminished for some subgroups (forest plots: supplementary fig A; τ^2, R, and prediction intervals: supplementary table C). For cut-off value ≥10, τ^2 values ranged from 0 to 6.97 for sensitivity and 0 to 1.65 for specificity for semistructured interviews, 0.32 to 1.28 for sensitivity and 0.32 to 1.48 for specificity for fully structured interviews (MINI excluded), and 0.21 to 1.44 for sensitivity and 0.07 to 0.71 for the MINI. The 95% prediction intervals for sensitivity and specificity were much wider than the corresponding 95% confidence intervals in supplementary table E, similarly reflecting the moderate to high heterogeneity between studies.

Nomograms for positive and negative predictive values of the PHQ-9 for hypothetical major depression

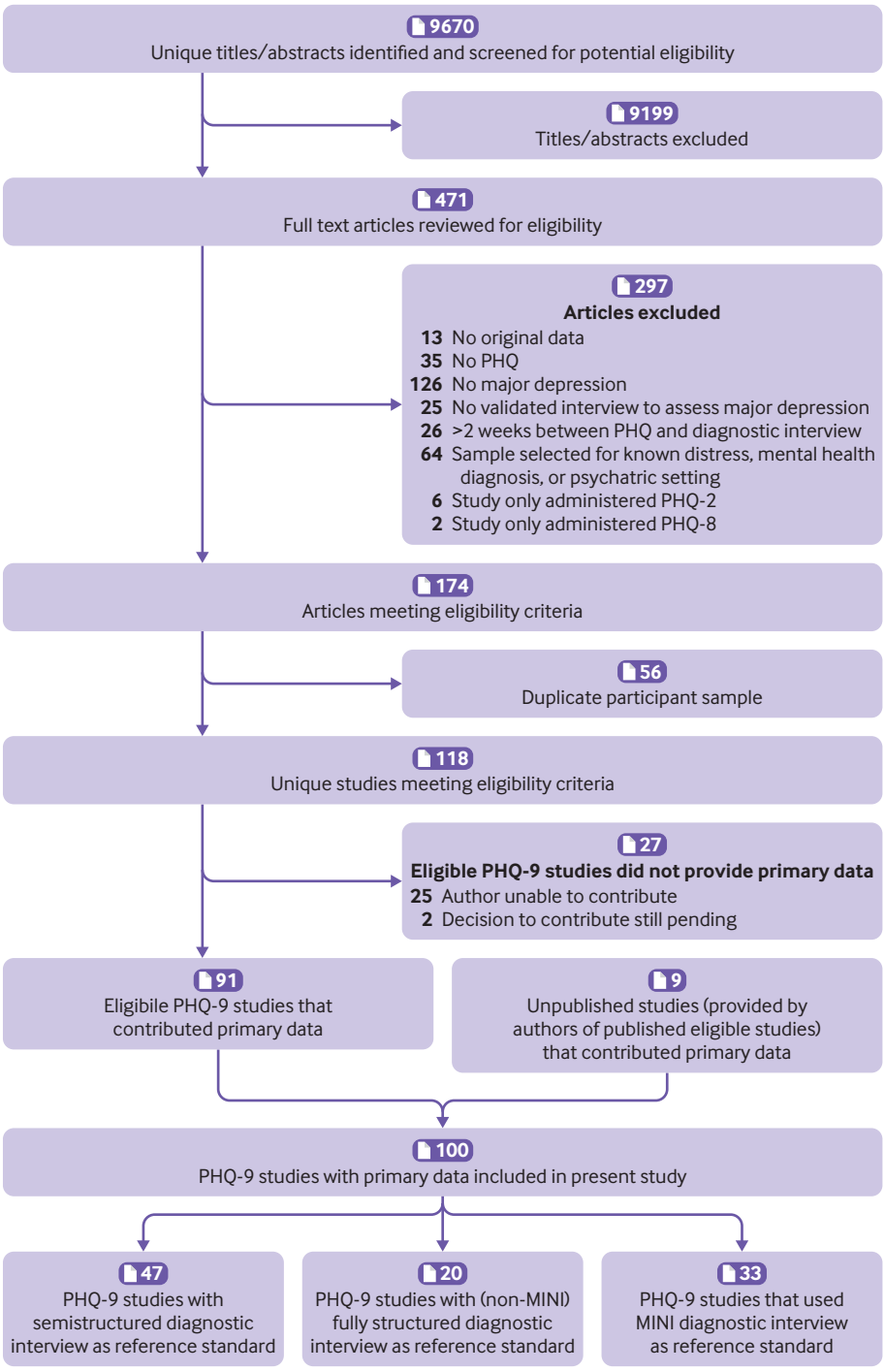

Fig 1 | Flow diagram of study selection process. MINI=Mini International Neuropsychiatric Interview; PHQ=Patient Health Questionnaire

prevalence values of 5-25% at a cut-off score of ≥10 are presented in figure 3. For these theoretical prevalence values, positive predictive values ranged from 23% to 65% for semistructured interviews, 22% to 64% for fully structured interviews, and 26% to 69% for the MINI; corresponding negative predictive values were 94% to 99%, 88% to 98%, and 91% to 99%, respectively.

Multiple meta-regression results showing the association between PHQ-9 accuracy and the three reference standard categories are shown in supplementary table D. Significant associations were found between reference standards and PHQ-9 sensitivity across cut-off values of 5-15. Compared with semistructured interviews, sensitivity was 5-23% (median 19%) higher than with fully structured interviews and 1-15% (median 10%) higher than with the MINI. Across cut-off values, the magnitude of estimated differences based on meta-regression were within 2-3% of difference estimates based on the

Table 1 | Distribution of participant data by diagnostic interview

Diagnostic interview	Studies (No)	Participants (No)	Major depression (No (%))
Semistructured:	47	11 234	1528 (14)
Structured Clinical Interview for DSM	44	9242	1389 (15)
Schedules for Clinical Assessment in Neuropsychiatry	2	1892	130 (7)
Depression Interview and Structured Hamilton	1	100	9 (9)
Fully structured:	20	17 167	1352 (8)
Composite International Diagnostic Interview	17	15 759	1067 (7)
Diagnostic Interview Schedule	1	1006	221 (22)
Clinical Interview Schedule-revised	2	402	64 (16)
Mini International Neuropsychiatric Interview	33	16 102	1661 (10)
Total	100	44 503	4541 (10)

DSM=Diagnostic and Statistical Manual.

bivariate random effects model. The maximum number of participants excluded from any subgroup analysis because a study had no patients with or without major depression was 63.

Among the 27 studies that did not contribute individual participant data, 13 published eligible accuracy data (supplementary table B2). Six used semistructured interviews, two fully structured interviews, and five the MINI, although two of the semistructured studies were excluded from the sensitivity analysis because they did not publish the number of participants with or without major depressive disorder. Including published results from the remaining 11 studies did not change results as shown in supplementary tables D14 to D16.

PHQ-9 accuracy among participants not diagnosed or receiving treatment for a mental health problem compared with all participants

Sensitivity estimates for participants not diagnosed or receiving treatment were not significantly different than those for all participants for any reference standard category. Specificity estimates were statistically significantly different for semistructured and MINI interviews. Among participants not currently diagnosed or receiving treatment compared with all participants, specificity was 1-4% (median 4%) higher across cut-off values with semistructured interviews and 1-6% (median 3%) higher across cut-off values with the MINI (supplementary table E). Specificity was not significantly different for fully structured interviews.

PHQ-9 sensitivity and specificity among subgroups and risk of bias

Meta-regression results interacting PHQ-9 sensitivity and specificity with reference standards and then with other subgroup variables stratified within reference standards are shown in supplementary table D. Pooled sensitivity and specificity and associated 95% confidence intervals for cut-off values 5-15 by reference standard and subgrouping variables are presented in supplementary table E. Receiver operating characteristic curves and corresponding areas under the curve are shown in supplementary figure B.

Table 2 | Distribution of participant data by subgroup*

	Semistructured diagnostic interviews			Fully structured diagnostic interviews			Mini International Neuropsychiatric Interview		
Participant subgroup	Studies (No)	Participants (No)	Major depression No (%)	Studies (No)	Participants (No)	Major depression No (%)	Studies (No)	Participants (No)	Major depression No (%)
All participants	47	11 234	1528 (14)	20	17 167	1352 (8)	33	16 102	1661 (10)
Participants not currently diagnosed or receiving treatment for a mental health problem	26	3687	603 (16)	5	4001	289 (7)	15	8365	578 (7)
Age <60 years	42	7349	1131 (15)	20	13 784	1087 (8)	31	10 489	1119 (11)
Age ≥60 years	39	3860	397 (10)	15	3374	265 (8)	27	5585	533 (10)
Women	46	6986	1040 (15)	20	9603	793 (8)	32	9574	1126 (12)
Men	39	4168	488 (12)	18	7554	557 (7)	30	6511	534 (8)
Country with very high human development index	38	9156	1047 (11)	16	15 422	1149 (7)	21	10 484	1108 (11)
Country with high human development index	5	811	215 (27)	0	0	0	7	3753	237 (6)
Country with low or medium human development index	4	1267	266 (21)	4	1745	203 (12)	5	1865	316 (17)
Non-medical care	2	567	105 (19)	4	8219	371 (5)	9	7802	117 (15)
Primary care	14	4566	683 (15)	7	4746	425 (9)	9	5063	543 (11)
Inpatient specialty care	12	2355	257 (11)	2	593	72 (12)	3	473	106 (22)
Outpatient specialty care	21	3746	483 (13)	7	3609	484 (13)	12	2634	511 (19)

*Some variables were coded at the study level, while others were coded at the participant level. Thus, the number of studies does not always add up to the total for each reference standard.

Table 3 | Comparison of sensitivity (95% confidence interval) and specificity (95% confidence interval) estimates among semistructured, full structured, and MINI reference standards

	Semi structured reference standard*		Fully structured reference standard†		MINI reference standard‡	
Cut-off score	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
5	0.98 (0.95 to 0.99)	0.53 (0.49 to 0.58)	0.91 (0.85 to 0.95)	0.61 (0.51 to 0.69)	0.96 (0.93 to 0.97)	0.60 (0.55 to 0.64)
6	0.97 (0.94 to 0.98)	0.61 (0.57 to 0.65)	0.88 (0.80 to 0.93)	0.69 (0.60 to 0.76)	0.92 (0.89 to 0.95)	0.68 (0.63 to 0.72)
7	0.95 (0.92 to 0.98)	0.68 (0.64 to 0.72)	0.82 (0.73 to 0.89)	0.75 (0.67 to 0.82)	0.88 (0.83 to 0.92)	0.74 (0.70 to 0.78)
8	0.92 (0.88 to 0.95)	0.74 (0.70 to 0.77)	0.77 (0.66 to 0.86)	0.81 (0.74 to 0.86)	0.85 (0.79 to 0.89)	0.80 (0.76 to 0.83)
9	0.89 (0.84 to 0.92)	0.80 (0.76 to 0.82)	0.69 (0.59 to 0.78)	0.85 (0.79 to 0.90)	0.80 (0.73 to 0.85)	0.85 (0.82 to 0.88)
10	0.85 (0.79 to 0.89)	0.85 (0.82 to 0.87)	0.64 (0.53 to 0.74)	0.88 (0.83 to 0.92)	0.74 (0.67 to 0.79)	0.89 (0.86 to 0.91)
11	0.81 (0.75 to 0.86)	0.88 (0.85 to 0.90)	0.57 (0.46 to 0.67)	0.91 (0.87 to 0.94)	0.67 (0.60 to 0.73)	0.91 (0.89 to 0.93)
12	0.75 (0.69 to 0.80)	0.90 (0.88 to 0.92)	0.52 (0.41 to 0.63)	0.93 (0.89 to 0.95)	0.61 (0.54 to 0.68)	0.93 (0.91 to 0.95)
13	0.67 (0.61 to 0.72)	0.93 (0.91 to 0.94)	0.45 (0.35 to 0.56)	0.95 (0.92 to 0.97)	0.55 (0.47 to 0.62)	0.95 (0.93 to 0.96)
14	0.61 (0.55 to 0.67)	0.94 (0.93 to 0.96)	0.39 (0.30 to 0.50)	0.96 (0.94 to 0.97)	0.47 (0.41 to 0.54)	0.96 (0.95 to 0.97)
15	0.52 (0.46 to 0.58)	0.96 (0.94 to 0.97)	0.32 (0.24 to 0.41)	0.97 (0.95 to 0.98)	0.40 (0.35 to 0.46)	0.97 (0.96 to 0.98)

MINI=Mini International Neuropsychiatric Interview.
*Number of studies=47; number of participants=11 234; number of participants with major depression=1528.
†Number of studies=20; number of participants=17 167; number of participants with major depression=1352.
‡Number of studies=33; number of participants=16 102; number of participants with major depression=1661.

The age and sex of participants, but no other variables, were statistically significantly associated with specificity in all three reference standard categories (supplementary table D). Specificity was higher for participants aged 60 or older than for younger participants. It was 2-12% (median 6%) higher with semistructured interviews, 3-11% (median 6%) higher for fully structured interviews, and 0-8% (median 2%) higher for the MINI; across reference standards, it was 0-12% (median 5%). Specificity was also higher for men than women. Differences from the meta-regression were 1-10% (median 4%), 1-7% (median 3%), and 0-7% (median 3%) for semistructured, fully structured, and the MINI reference standards, respectively; across reference standards, it was 0-10% (median 3%). For age and sex, magnitudes of differences based on meta-regression were within 1-5% of difference estimates based on the bivariate random effects meta-analytic model.

Based on results from studies with semistructured interviews, the cut-off value that maximised combined sensitivity and specificity shifted slightly from ≥10 for some subgroups. By age, the maximum was obtained with a cut-off value of ≥11 for age less than 60 and ≥10 for age 60 or older. By sex, it was ≥11 for women and ≥9 for men. However, in all instances, these maximum values were within 1-2% of those obtained with a cut-off value of ≥10. Results were similar for fully structured interviews and the MINI (supplementary table E).

QUADAS-2 ratings for all included primary studies are presented in supplementary table F. Of 395 study level items, 12 were rated as high, 130 as unclear, and 253 as low risk of bias. No QUADAS-2 signalling questions were consistently associated with PHQ-9 sensitivity or specificity, as shown in supplementary table D.

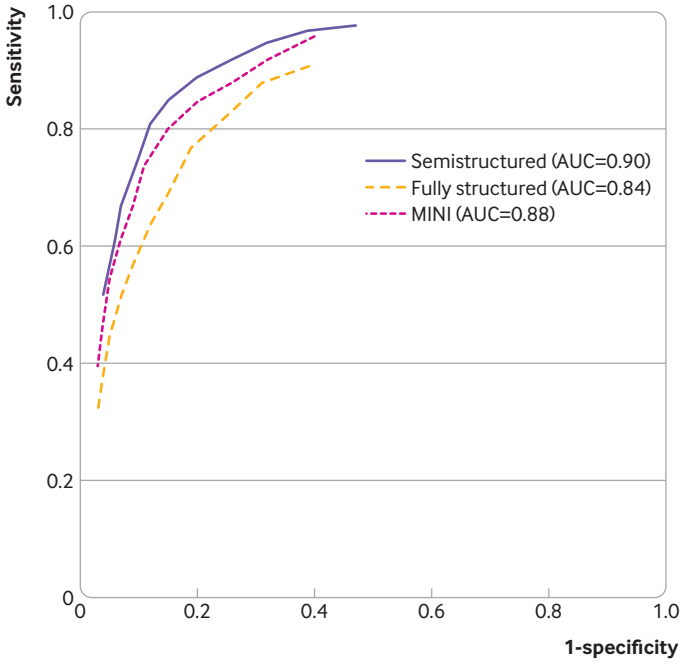

Fig 2 | Receiver operating characteristic curves showing estimates of sensitivity and specificity at each cut-off value and each reference standard category. AUC=area under the curve; MINI=Mini International Neuropsychiatric Interview

Discussion

Principal findings

We evaluated PHQ-9 accuracy for screening for major depression. We found that combined PHQ-9 sensitivity (85%) and specificity (85%) was maximised at the standard cut-off value of ≥10 among studies using a semistructured interview, which is the interview type designed to replicate diagnostic procedures most closely. When only participants not already diagnosed or receiving treatment were considered, which reflects the population that would be screened in practice, sensitivity was unchanged at a cut-off value ≥10, whereas specificity improved to 89%.

Age and sex were statistically significantly associated with PHQ-9 specificity in all three types of diagnostic interviews. The PHQ-9 was more specific for participants aged 60 or older than for younger participants and for men than for women. Sensitivity was not associated with age or sex. Differences in accuracy by subgroups resulted in different cut-off values that maximised combined sensitivity and

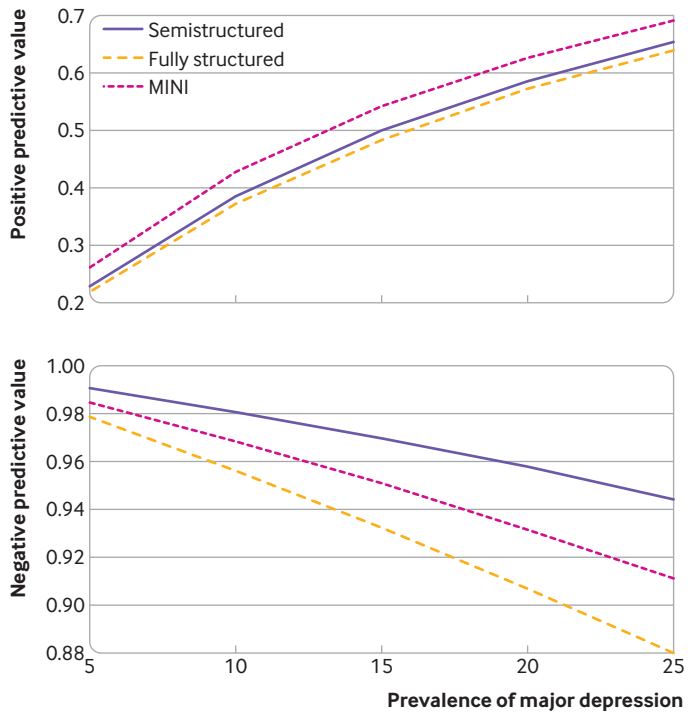

Fig 3 | Nomograms of positive and negative predictive values for cut-off score 10 of the Patient Health Questionnaire-9 for major depression prevalence values of 5-25% for semistructured, fully structured, and MINI diagnostic interviews. MINI=Mini International Neuropsychiatric Interview

specificity in some subgroups, but the margin of difference compared with the standard cut-off value of ≥10 was minimal in all instances and not large enough to warrant the use in practice of different cut-off values for patients with different demographic characteristics.

Comparison with other studies

This IPDMA included data from almost twice as many primary studies and from approximately two and a half times as many participants as the previous PHQ-9 IPDMA.[25] Many results were similar. In both studies, sensitivity was substantially higher in comparison with a semistructured reference standard than for fully structured and MINI reference standards. This finding is consistent with findings that compared with semistructured interviews, fully structured and MINI reference standards generate a substantial number of false positive diagnoses, controlling for participant and study characteristics[26-29]; the additional diagnoses that are generated would be expected to result in lower sensitivity of the PHQ-9 to detect people with major depression, which we found to be the case. Sensitivity and specificity were maximised at a cut-off score of ≥10 among studies using a semistructured interview. Older age was associated with significantly, although minimally, greater specificity across reference standards for most cut-off values, a finding which contradicts assumptions that screening tools might be less accurate in elderly people and suggests that the PHQ-9 is similarly or more accurate.

In contrast to the previous IPDMA, we found that PHQ-9 specificity might be higher for men than for women across reference standards. We also found that specificity was significantly higher when only data from participants not diagnosed or receiving treatment for a mental health problem were examined among studies using a semistructured or the MINI reference standard; based on studies with semistructured interviews, it was approximately four percentage points higher. Previous studies[51 52] have predicted that inclusion of such participants might bias accuracy results, although those studies suggested that the bias would probably be through improved sensitivity. Instead, we found that specificity was reduced when people who might have already been diagnosed or in treatment were considered, suggesting that false positive screens occurred in this group. Magnitude of bias was small, however, and this result was not seen among studies that administered fully structured interviews.

Implications

Many studies report the cut-off value that maximises combined sensitivity and specificity, and we did this in the present study to establish a reference point. There is no clinical reason, however, for selecting a cut-off value based on this standard for use in clinical trials or practice. Higher cut-off values would rule out more participants without depression but would detect fewer participants who meet the criteria for major depression. Conversely, lower cut-off values would detect more participants who met diagnostic criteria at the expense of more false positive screens among people without major depression. Ideally, clinical decision making would consider the benefits versus the costs and harms that are generated from correct and incorrect screening results and the expected net benefit or costs and harms at all possible cut-off values.[62] Selecting a cut-off value in this way depends on local values and resources, as well as assumptions about outcomes from positive and negative screening tests at different cut-off values.

Ideally, clinical trials could inform researchers and practitioners about outcomes from using different cut-off values. To the best of our knowledge, however, only one depression screening trial[59] has randomised participants to screening and no-screening groups and used the PHQ-9 or the eight item PHQ-8, which performs virtually identically,[40] to identify participants with possible depression. Kronish et al[63] used a cut-off value of ≥10 on the PHQ-8 to assign participants with an acute coronary syndrome in the past 12 months to notification of primary care clinicians of screening results, notification of clinicians plus stepped depression care, or usual care with no notification. People already receiving depression treatment were excluded. Only 7% of participants, however, had positive screening tests at the standard cut-off value of ≥10; fewer than 40 patients with positive screens were found in each intervention arm, some of whom might not have completed offered interventions, and results did not show benefit from screening. The small number of positively screened participants who could have been offered an intervention, however, makes it difficult to draw conclusions about the

doi: 10.1136/bmj.n2183 | *BMJ* 2021;374:n2183 | the bmj

appropriateness of the cut-off value. One might suggest that a lower cut-off value could have been used to increase positive screens and, potentially, patients with depression, but this would also have resulted in more false positive screens and a heavier assessment burden as the resources consumed by unnecessary assessment (typically by a mental health professional, depending on the setting) divert scarce resources that might otherwise be used for intervention. Furthermore, it is possible that if the number of true positive screens had been increased, those patients might have been people with mild symptoms who are less likely to benefit from intervention. Another possibility would be to increase the cut-off threshold, which would require more participants to be screened in a trial; an obvious disadvantage would be the sheer number of people who would have to be screened to identify those eligible for intervention.

Thus there are no easy answers for selection of the most appropriate cut-off value, and researchers and clinicians who wish to screen with the PHQ-9 will need to examine its likely performance at different cut-off values. To help clinicians do this, we have created a web-based tool (depressionscreening100.com/phq) based on this study's findings. The tool estimates the expected numbers of positive and negative screens and true and false screening results based on different assumed prevalence and different cut-off values (box 1).

Strengths and limitations

Strengths of our IPMDA method for evaluating PHQ-9 accuracy, compared with conventional meta-analyses, include (a) integration of data from studies that collected PHQ-9 and reference standard data but did not publish accuracy results; (b) inclusion of data from eligible participants in studies that included some eligible and some ineligible participants (eg, those already receiving mental healthcare) by selecting eligible participants only; (c) inclusion of studies that published results based on composite reference standards (eg, any psychiatric disorder) by coding based on the presence of major depression; (d) the ability to conduct subgroup analyses by participant or study characteristics; few primary studies have attempted subgroup analyses because of the amount of data required; and (e) the ability to conduct analyses for all relevant cut-off values for all included studies and reduce bias from selective cut-off value reporting, which occurs because many studies report results for only some cut-off values, often those with the most positive results.[64 65] The present IPDMA included data from 44 503 participants an increase of 27 146 from our previous IPDMA (N=17 357).[25]

This study has some limitations. Firstly, we could not include data from 27 of 127 eligible studies (21%; 14% of eligible participants), although including data from eligible studies that published sensitivity and specificity but did not contribute data did not change the IPDMA results. Secondly, there was substantial heterogeneity, although this was reduced somewhat in subgroup analyses. Methods for estimating and interpreting heterogeneity in meta-analyses of test accuracy are not well established, and there are no established guidelines for interpretating results of the quantitative metrics we used. High heterogeneity in meta-analyses of test accuracy studies is common.[37 57] Thirdly, because 41% of participants had missing medical comorbidity data, and most languages and countries were represented by few studies, we could not conduct those subgroup analyses. Fourthly, primary studies were classified according to the diagnostic interview used, but interviewers might not have always administered the interviews as intended, which could have influenced results. Fifthly, because of the time required to determine whether datasets used in published studies are eligible; to invite authors to participate; to arrange for data transfer, including data transfer agreements; and to conduct quality control and data harmonisation procedures, studies included in the IPDMA were published up to May 2018, and more recent studies could not be included.

Conclusions

We found that PHQ-9 sensitivity and specificity were both 85% compared with semistructured interviews. Sensitivity was unchanged, but specificity was higher (89%) when only people eligible for screening in practice were considered. Specificity appears to be higher for participants aged 60 or older and for men, but differences are not large enough to consider subgroup specific cut-off value thresholds. Clinicians who use the PHQ-9 to screen should select a cut-off point that provides the best balance of their preferences and resources for sensitivity and specificity and true and false positive screens.

Box 1: Putting results into practice

- Standard cut-off points that are commonly used for depression screening, including the cut-off value of ≥10 on the Patient Health Questionnaire-9, are typically selected because they maximise combined sensitivity and specificity
- Maximising sensitivity and specificity, however, does not necessarily maximise the likelihood of patient benefits, minimise costs and harms, or reflect local concerns, such as capacity for conducting assessments of people with positive screens
- Researchers and clinicians can choose a cut-off value based on clinical priorities and local resources by comparing screening outcomes that would occur with different outcomes, including true and false positive screens and true and false negative results
- A knowledge translation tool (www.depressionscreening100.com/phq) based on the findings from this study can be used to generate screening outcomes for different cut-off values based on local assumptions about prevalence

AUTHOR AFFILIATIONS

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

[2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, WC, Canada

[3]Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK

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

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

[6]Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, 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

Contributors: ZFN, BLevis, ABenedetti, and BDT were responsible for the study conception and design. ZFN, BLevis, YS, CH, AK, YW, PMB, DN, EB, and BDT contributed to data extraction, coding, evaluation of included studies, and data synthesis. ZFN, BLevis, ABenedetti, and BDT contributed to data analysis and interpretation. ZFN, BLevis, ABenedetti, 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 PHQ Group contributed to data extraction, coding, and synthesis (MImran, DBR, KER, MA, AWL); via the design and conduct of database searches (JTB, LAK); as members of the DEPRESSD steering committee, including conception and oversight of collaboration (PC, SG, JPAI, DM, SBP, IS, RCZ); as a knowledge user consultant (SM); by contributing included datasets (DHA, SHA, DA, BA, LA, HRB, ABeraldi, CNB, ABhana, CHB, RIB, PB, GC, MHC, JCNC, LFC, DC, RC, NC, KC, AC, YC, FMD, JMdMvG, JD, CDQ, JRF, FHF, SF, JRWF, DF, ECG, BG, LG, LJG, FGS, EPG, CGG, BJH, LHantsoo, EEH, MHärter, UH, LHides, SHE, SH, MHudson, TH, MInagaki, KI, HJJ, NJ, MEK, KMK, SK, BAK, YK, FL, MAL, HFLA, SIL, ML, SRL, BLöwe, NPL, CL, RAM, LM, BPM, AM, SMS, TNM, KM, JEMN, LN, FLO, VP, BWP, PP, IP, AP, SLP, TJQ, ER, SDR, KR, AGR, HJR, ISS, MTS, JS, EHS, ASidebottom, ASimning, LSpangenberg, LStafford, SCS, KS, PLLT, MTR, TDT, AT, CMvdFC, TvH, HCvW, PAV, LIW, JLW, WW, DW, JW, MAW, KWinkley, KWynter, MY, QZZ, YZ). All authors, including group authors, provided a critical review and approved the final manuscript.

Funding: This study was funded by the Canadian Institutes of Health Research (CIHR; KRS-134297, PCG-155468, PJT-162206). ZFN was supported by the Mitacs Accelerate Postdoctoral Fellowship. BLevis and YW were supported by Fonds de recherche du Québec - Santé (FRQS) Postdoctoral Training Fellowships. PMB was supported by a studentship from the Research Institute of the McGill University Health Centre. DN was supported by GR Caverhill Fellowship from the Faculty of Medicine, McGill University. ABenedetti was supported by a FRQS researcher salary award. BDT was supported by a Tier 1 Canada Research Chair. DBR was supported by a Vanier Canada Graduate Scholarship. 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 ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: support from the Canadian Institutes of Health Research 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: Dr Bernstein has consulted to Abbvie Canada, Amgen Canada, Bristol Myers Squibb Canada, Roche Canada, Janssen Canada, Pfizer Canada, Sandoz Canada, Takeda Canada, and Mylan Pharmaceuticals; 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. Dr Chan J CN is a steering committee member and/or consultant of Astra Zeneca, Bayer, Lilly, MSD and Pfizer. She has received sponsorships and honorarium for giving lectures and providing consultancy and her affiliated institution has received research grants from these companies. Dr Chan LF declares personal fees and non-financial support from Otsuka, Lundbeck, and Johnson and Johnson; and non-financial support from Ortho-McNeil-Janssen, and Menarini, outside the submitted work. Dr Hegerl declares that within the last three years, he was an advisory board member for Lundbeck and Servier; a consultant for Bayer Pharma; a speaker for Pharma and Servier; and received personal fees from Janssen Janssen and a research grant from Medice, all outside the submitted work. Dr Inagaki declares that he has received personal fees from Meiji, Mochida, Takeda, Novartis, Yoshitomi, Pfizer, Eisai, Otsuka, MSD, Sumitomo Dainippon, Janssen, and Eli Lilly, all outside of the submitted work. Dr Pugh declares that she received salary support from Pfizer-Astella and Millennium, outside the submitted work. Dr Rancans declares that he received grants, personal fees and non-financial support from Gedeon Richter; personal fees and non-financial support from Lundbeck, Servier, and Janssen Cilag; personal fees from Zentiva, and Abbvie; outside the submitted work. Dr Schram declares that the data collection of the primary study by Janssen et al. was supported by unrestricted grants from Janssen, Novo Nordisk, and Sanofi. Dr Wagner declares that she receives personal fees from Celgene, outside the submitted work. 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: The research ethics committee of the Jewish General Hospital declared that research ethics approval was not required, since the study involved IPDMA of anonymised previously collected data. However, for each included dataset, we confirmed that the original study received ethics approval and that participants provided informed consent.

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

The manuscript's guarantor affirms that this 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 study participants and related patient and patient communities: There are no plans to disseminate the results of the research to study participants or the relevant patient community. However, a web based knowledge translation tool, intended for clinicians (the end users of the PHQ-9 screening tool), is available at depressionscreening100.com/phq. 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 Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet* 2007;370:851-8. doi:10.1016/S0140-6736(07)61415-9
2 Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet* 2006;367:1747-57. doi:10.1016/S0140-6736(06)68770-9
3 Mathers CD, Lopez AD, Murray CJL. The burden of disease and mortality by condition: Data, methods, and results for 2001. In: Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJL, eds. *Global Burden of Disease and Risk Factors.* The International Bank for Reconstruction and Development/The World Bank Group, 2006:45-93.
4 Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* 2013;382:1575-86. doi:10.1016/S0140-6736(13)61611-6
5 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
6 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
7 Allaby M. *Screening for depression: A report for the UK National Screening Committee (Revised report).* UK National Screening Committee, 2010.
8 Thombs BD, Markham S, Rice DB, Ziegelstein RC. Does depression screening in primary care improve mental health outcomes?*BMJ* 2021;374:n1661. doi:10.1136/bmj.n1661
9 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
10 Gilbody S, Sheldon T, Wessely S. Should we screen for depression?*BMJ* 2006;332:1027-30. doi:10.1136/bmj.332.7548.1027
11 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

12 Thombs BD, Ziegelstein RC. Does depression screening improve depression outcomes in primary care?*BMJ* 2014;348:g1253. doi:10.1136/bmj.g1253
13 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
14 Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606-13. doi:10.1046/j.1525-1497.2001.016009606.x
15 Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann* 2002;32:1-7. doi:10.3928/0048-5713-20020901-06
16 Spitzer RL, Kroenke K, Williams JB, Primary Care Evaluation of Mental Disorders. 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. doi:10.1001/jama.282.18.1737
17 Maurer DM, Raymond TJ, Davis BN. Depression: Screening and Diagnosis. *Am Fam Physician* 2018;98:508-15.
18 American Academy of Family Physicians. Clinical preventive service recommendation. Depression. https://www.aafp.org/family-physician/patient-care/clinical-recommendations/all-clinical-recommendations/depression.html.
19 *Diagnostic and statistical manual of mental disorders: DSM-III.* 3rd ed, revised. American Psychiatric Association, 1987.
20 *Diagnostic and statistical manual of mental disorders: DSM-IV.* 4th ed. American Psychiatric Association, 1994.
21 *Diagnostic and statistical manual of mental disorders: DSM-IV.* 4th ed, text revised. American Psychiatric Association, 2000.
22 *Diagnostic and statistical manual of mental disorders: DSM-V.* 5th ed. American Psychiatric Association, 2013.
23 Wittkampf KA, Naeije L, Schene AH, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen Hosp Psychiatry* 2007;29:388-95. doi:10.1016/j.genhosppsych.2007.06.004
24 Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* 2007;22:1596-602. doi:10.1007/s11606-007-0333-y
25 Levis B, Benedetti A, Thombs BD, DEPRESsion 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
26 Wu Y, Levis B, Ioannidis JPA, Benedetti A, Thombs BD, DEPRESsion 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
27 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
28 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
29 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
30 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
31 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
32 Nosen E, Woody SR. Diagnostic Assessment in Research. In: McKay, D . *Handbook of research methods in abnormal and clinical psychology.* Sage, 2008:109-24.
33 Kurdyak PA, Gnam WH. Small signal, big noise: performance of the CIDI depression module. *Can J Psychiatry* 2005;50:851-6. doi:10.1177/070674370505001308
34 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
35 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
36 Thombs BD, Benedetti A, Kloda LA, et al. The diagnostic accuracy of the Patient Health Questionnaire-2 (PHQ-2), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-9 (PHQ-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *Syst Rev* 2014;3:124. doi:10.1186/2046-4053-3-124
37 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
38 Stewart LA, Clarke M, Rovers MP, et al, RISMA-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
39 Levis B, Sun Y, He C, et al, Depression Screening Data (DEPRESSD) PHQ Collaboration. Accuracy of the PHQ-2 Alone and in Combination With the PHQ-9 for Screening to Detect Major Depression: Systematic Review and Meta-analysis. *JAMA* 2020;323:2290-300. doi:10.1001/jama.2020.6504
40 Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychol Med* 2020;50:1368-80. doi:10.1017/S0033291719001314
41 The ICD-10 Classifications of Mental and Behavioural Disorder. *Clinical Descriptions and Diagnostic Guidelines Geneva.* World Health Organization, 1992.
42 PRESS – Peer Review of Electronic Search Strategies: 2015 Guideline Explanation and Elaboration (PRESS E&E). Ottawa: CADTH; 2016 Jan.
43 United Nations. International Human Development Indicators. http://hdr.undp.org/en/countries.
44 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
45 First MB. *Structured clinical interview for the DSM (SCID).* John Wiley & Sons, Inc, 1995.
46 World Health Organization. *Schedules for clinical assessment in neuropsychiatry: manual.* Amer Psychiatric Pub Inc, 1994.
47 Freedland KE, Skala JA, Carney RM, et al. The Depression Interview and Structured Hamilton (DISH): rationale, development, characteristics, and clinical validity. *Psychosom Med* 2002;64:897-905.
48 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
49 Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychol Med* 1992;22:465-86. doi:10.1017/S0033291700030415
50 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
51 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
52 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
53 van der Leeden R, Busing FMTA, Meijer E. *Bootstrap methods for two-level models. Technical Report PRM 97-04.* Leiden University, Department of Psychology, 1997.
54 van der Leeden R, Meijer E, Busing FMTA. Resampling multilevel models. In: Leeuw J, Meijer E, eds. *Handbook of multilevel analysis New York.* Springer, 2008: 401-33. doi:10.1007/978-0-387-73186-5_11
55 Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials* 2010;11:85. doi:10.1186/1745-6215-11-85
56 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

57 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.
58 Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539-58. doi:10.1002/sim.1186
59 R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
60 RStudio Team. (2020). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL https://www.rstudio.com/.
61 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
62 Smits N, Smit F, Cuijpers P, De Graaf R. Using decision theory to derive optimal cut-off scores of screening instruments: an illustration explicating costs and benefits of mental health screening. *Int J Methods Psychiatr Res* 2007;16:219-29. doi:10.1002/mpr.230
63 Kronish IM, Moise N, Cheung YK, et al. Effect of depression screening after acute coronary syndromes on quality of life: the CODIACS-QoL randomized clinical trial. *JAMA Intern Med* 2020;180:45-53. doi:10.1001/jamainternmed.2019.4518
64 Levis B, Benedetti A, Levis AW. Selective cutoff reporting in studies of diagnostic test accuracy: a comparison of conventional and individual-patient-data meta-analysis of the Patient Health Questionnaire-9 depression screening tool. *Am J Epidemiol* 2017;185:954-64.
65 Neupane D, Levis B, Bhandari PM, et al . Selective cutoff reporting in studies of the accuracy of the Patient Health Questionnaire-9 and Edinburgh Postnatal Depression Scale: Comparison of results based on published cutoffs versus all cutoffs using individual participant data meta-analysis. *Int J Methods Psychiatr Res* 2021;30:e1873.

Web appendix 1: Supplementary material
Web appendix 2: Depression Screening Data (DEPRESSD) PHQ Group members
Web appendix 3: Funding of primary studies